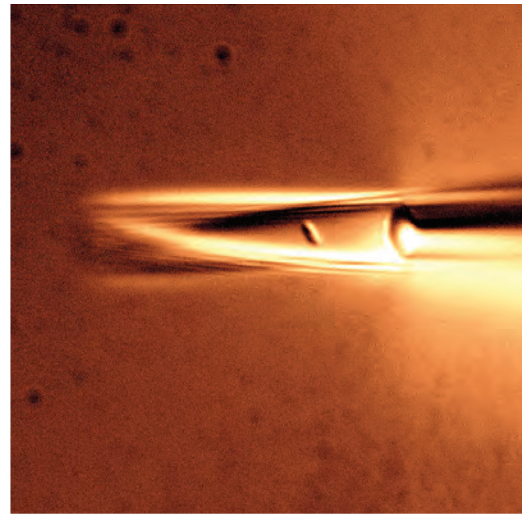
U.S. Department of Energy
Joint Genome Institute (JGI)

A 10-Year
Strategic Vision

Draft, October, 2011



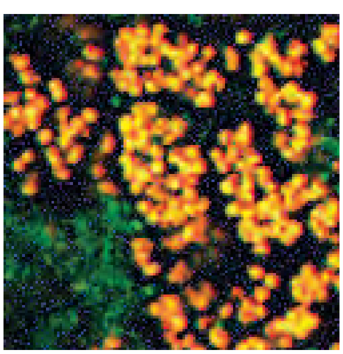
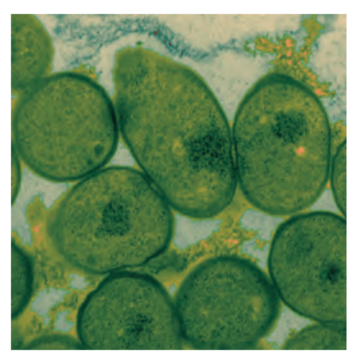
FORGING THE FUTURE OF THE DOE JGI

U.S. DEPARTMENT OF ENERGY
Office of Science

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

OUR VISION

The user facility pioneering functional genomics to solve the most relevant bioenergy and environmental problems

U.S. Department of Energy
Joint Genome Institute (JGI)

A 10-Year
Strategic Vision

FORGING THE FUTURE OF THE DOE JGI

Draft, October, 2011

This document is broken into three sections:

I. Introduction

II. Background-Science Drivers

III. Capabilities

The Introduction provides a high level overview of the DOE Joint Genome Institute (DOE JGI) and how it plans to evolve as a genomic user facility to meet the scientific needs of energy and environmental research over the next decade. The Background-Science Driver section provides an assessment of the major scientific energy and environmental problems that the DOE JGI needs to enable its users to solve. Finally the Capabilities component is broken into three sections, "pillars", which outline the capabilities and operating principles of the JGI as it transitions into becoming a next-generation genome center.

TABLE OF CONTENTS

I. INTRODUCTION

EXECUTIVE SUMMARY

An extraordinary biological revolution is taking place, spurred by advances in molecular biology and genome-scale approaches. The promise of this revolution is enormous. As we learn more about the basic biological mechanisms involved in energy generation, environmental processes, and climate systems, the opportunities to turn this knowledge to our benefit continue to grow. A deep, systems-based understanding of natural biological processes will be critical for the U.S. Department of Energy mission to develop next-generation biofuels, advance global carbon management and successfully remediate contaminated environmental sites. Genome biology will play an essential role in these efforts, enabled by a technological revolution that is now in full stride. In this document, based on extensive input from DOE Joint Genome Institute (DOE JGI) users, advisory panels, several strategic planning meetings and workshops (see Appendix) and the 2010 DOE ***Grand Challenges for Biological and Environmental Research: A Long-Term Vision*** report, we aim to define the future role of the DOE JGI. We identify key strategic capabilities required for sustained scientific progress and leadership in energy and environmental genomics that will enable DOE JGI users to solve the most significant scientific challenges of the next decade.

DNA sequencing has become cheaper, faster and easier, but it represents only one of many genomic capabilities needed if these challenges are to be met. This document outlines a vision of the DOE JGI as a facility that will continue to generate sequence data, but that increasingly emphasizes complementary pre-sequencing, functional genomics and data analysis capabilities that will be required for carrying out the large-scale environmental and systems biological studies of the future. These changes in the mission focus and technological portfolio for the DOE JGI will be mirrored by its evolution from a production sequencing facility to a next-generation genome center, offering a diversity of capabilities that will complement massive-scale sequence production to meet the scientific needs of energy and environmental researchers. A partial listing of initial major areas of new development or expansion includes:

Large-scale rapid DNA synthesis and genomic manipulations. To accelerate the linking of sequence to function, the DOE JGI will develop new approaches for designing and creating DNA fragments encompassing genes and larger segments of DNA. These capabilities will be made available to users for testing of genomics-derived hypotheses, creation of synthetic pathways and for the functional exploration of metagenomic and other sequence data sets.

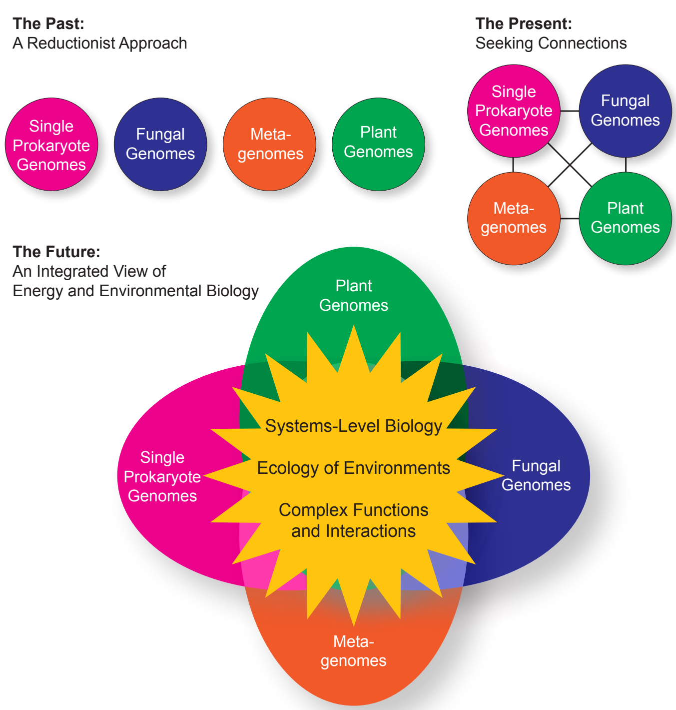
Massive-scale and customizable sample processing. The exponential growth in sequence data generation fortunately mirrors the expansive needs of future large-scale environmental and systems-based science. However, to be useful to researchers, this progress must be accompanied by parallel improvements in the scale of our ability to process samples for genomic characterization. The DOE JGI will develop custom large-scale sample-processing capabilities including the implementation of automated DNA/RNA extractions able to process tens of thousands of samples and large-scale single-cell and single-chromosome isolation techniques.

High performance computing infrastructure for data processing. Processing and integration of a rapidly increasing number and size of sequence datasets requires scalable methods running on a high-performance/high-throughput computing infrastructure. The DOE JGI will strengthen its strategic partnerships with supercomputing centers as providers of its computing needs and pursue collaborations with expert computer science and applied-mathematics groups to develop data processing and integration methods that are scalable and perform efficiently in a high-performance computing environment.

Multidimensional genome annotation and data integration. We will develop advanced data processing and integration techniques enabling data interpretation across the rapidly expanding universe of genomic, metagenomic and functional genomic datasets. This will include data fusion strategies that involve employing integration and reduction of multidimensional data to improve analytical accuracy. These capabilities will allow us to refine both structural annotations (the location of functional elements within sequences) and functional annotation (the function of these elements in the context of biological systems), raising the level of "interpreted" data provided to DOE JGI users.

Organization of mission-oriented user communities. As genome and functional genomic projects become larger and more complex, DOE JGI will play an expanded role in organizing communities around problems of central importance to DOE. DOE JGI will help to coordinate activities of diverse groups of scientists, ensure access to state-of-the-art genomics capabilities and strategies, and facilitate data sharing and integration in order to speed progress toward solving DOEs most pressing challenges in alternative fuels, carbon management and climate and environmental remediation.

Technological innovation in the field of genomics continually shifts and expands the questions that are feasible to address.



Integrated Biology will play a foundational role at the DOE JGI next-generation genome center.

The compartmentalization of the DOE JGI into four scientific programs (Plants, Microbes, Fungi, and Metagenomes) has been an important foundation of its organizational and scientific successes. However, many modern biological questions transcend this functional structure and DOE JGI's organization will adjust accordingly over time. The dynamic interactions between each of these domains in the environment are increasingly recognized and must be a focus of future biological studies, as outlined in the National Academy of Sciences (NAS) report, *A New Biology for the 21st Century*. As the structure of the DOE JGI evolves, integrative studies will blur the boundaries between domains and programs, enabling overarching and integrated research programs focused on the systems biology of environmental interactions.

Over the next decade, both predictable and unexpected technological innovations will continue to propel genomic science forward. Existing grand challenges will be completed; new grand challenges will emerge. The evolutionary strategy described in this document is intended to provide DOE JGI users with access to cutting-edge nucleic acid-based approaches as well as other synergistic capabilities, to reinforce the DOE JGI's unique position as the world's leading user facility in energy and environmental genomics.

EVOLUTION OF THE JGI AS A USER FACILITY

In the 12 years of its existence, the DOE JGI has transitioned through several phases as a user facility. The human genome era was the first phase. Here, the users were the large community of biologists requiring the specialized capabilities of the DOE JGI to sequence major portions of the human genome. The Institute at that time resembled, in some ways, a DOE accelerator facility for high-energy physics where all efforts focus on a single experiment agreed upon by a large community of scientists. In a second phase following the completion of the human genome, the DOE JGI morphed into a sequence-production user facility for energy and environmental genomics that more closely resembled a DOE synchrotron light source or a DOE supercomputer user facility. Individual and small groups of investigators with externally reviewed and prioritized projects were given access to the facility's unique capabilities to advance their science.

The recent development of a new generation of massively parallelized sequencing technologies has led to a widening gap between the rate of sequence generation and the ability to process and draw biological insights from the data.

EXAMPLES OF THE SCIENCE EXPECTED IN THE NEXT DECADE AS A RESULT OF USERS EXPLOITING DOE JGI GENOMICS CAPABILITIES:

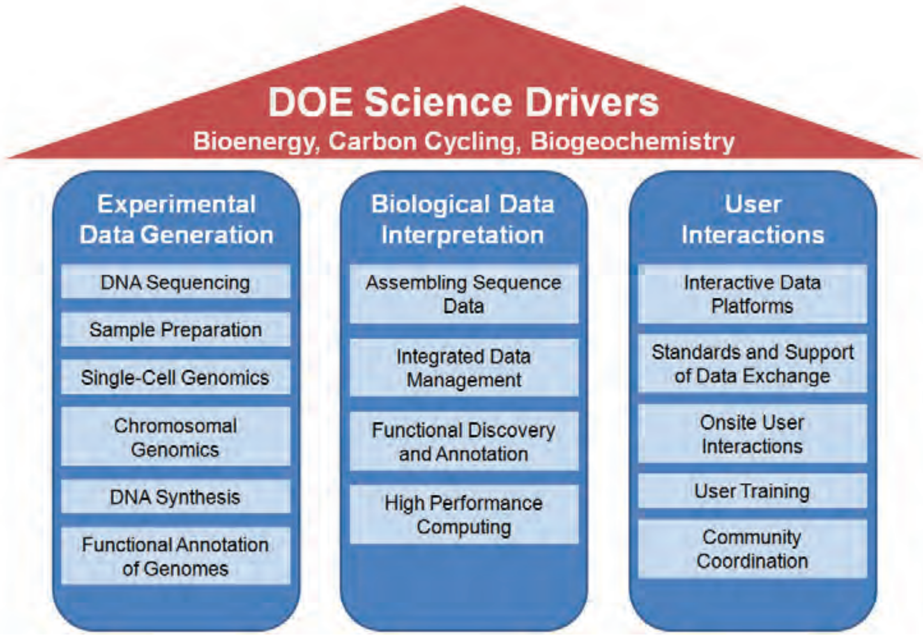
- Discover and study new branches of life and new metabolic activities through massive-scale isolation and sequencing of unexplored microbial and fungal "dark matter"
- Add functional annotation to all newly discovered gene families and predict the metabolic capabilities of all sequenced organisms and environments
- Real-time assembly and annotation of genomes from simple and complex environments
- Accurately predict the complete metabolome of single organisms and complex communities
- Enable users to develop biological reagents to mitigate environmental perturbations (e.g., oil spills) through the application of metagenome sequence analysis and organism engineering
- Enable users to engineer plants that recruit optimized rhizosphere communities through a systems-based understanding of plant-microbe interactions and biodesign capabilities
- Provide the data required for investigators to model and predict methane and carbon release from warming permafrost
- Generate the compendium of genomic information and capabilities that will enable scientists to bioengineer new organisms for direct conversion of biomass to biofuels

The generation of large amounts of sequence data is no longer a unique capability possessed by a limited number of large centralized facilities. Rather, the bottlenecks are now both upstream and downstream of sequencing. Upstream, there is a need for massive-scale and innovative sample processing, while downstream there is a need for more integrated informatics, as well as tighter linkage to functional studies. Many of the most important scientific challenges in energy and environmental research in the future will only be adequately tackled at large-scale facilities with multiple genomic capabilities and resources, including the support of multidisciplinary teams of specialized experts.

To respond to this challenge, the DOE JGI is once again evolving and this document marks the beginning of a third phase in the history of the DOE JGI as a user facility. The DOE JGI of the future will have expanded technologies and competencies for: processing samples, adding functional information to sequence data and integrating and displaying the data to accelerate the discovery of new insights from the data. In addition to large-scale DNA sequencing, multiple different complementary, state-of-the-art biological and computational capabilities will be made available to users. The DOE JGI will also increasingly work with and facilitate user access to complementary community-developed resources, both those generating data (e.g. EMSL for proteomic data) and those providing tools for data integration and analysis (e.g. the DOE Knowledgebase). The multiple cutting-edge experimental and computational technologies we envision establishing at the DOE JGI next-generation genome center will offer users access to capabilities that go far beyond those available in individual laboratories or core facilities, in both scale and diversity. Given the explosive advances occurring in genomics, we expect that the technologies available at the DOE JGI will evolve, but a constant will be the availability of a dedicated staff of scientists working with users to address important energy and environmental issues.

VISION OF A NEXT-GENERATION GENOME CENTER

We believe a useful metaphor for this strategic plan is to view the DOE JGI as a building (see page 4). The roof in this scheme would represent the most important scientific questions of energy and environmental research that will be explored in the next decade. This roof rests on pillars representing the large-scale genomic capabilities that will be required by scientists to address these questions. In the sections below we articulate the plan. In the "Background – Science Drivers" section, we highlight the major areas expected to represent the science drivers of the DOE JGI over the next decade. It is this science that dictates the capabilities that the DOE JGI will need to have. Following the Background section, we describe the Capabilities that will reside within the DOE JGI in the future. We have divided the needed capabilities into three general areas represented by pillars (1. Data Generation, 2. Biological Data Interpretation, 3. User Interactions). While these pillars are not meant to represent isolated activities, they do symbolize the major focus areas of the DOE JGI. Each of the three sections describing the individual pillars outlines a set of goals, stretch goals, and the science strategy that we envision will be required to achieve them. An overarching emphasis of the capabilities described in this strategic plan is to provide ways for users to link sequence to function targeted at contributing solutions for the most important DOE-relevant energy and environmental challenges.



Employing an evolving set of cutting-edge experimental and computational technologies, the DOE JGI next-generation genome center will empower users to perform studies at a scale and complexity far exceeding the capabilities of any individual laboratory.

UNIQUE CAPABILITIES TO BE MADE AVAILABLE TO USERS AT THE DOE JGI:

- Provide automated processing of tens of thousands of samples (plant or environmental) in a controlled reproducible manner for genomic and transcriptomic analysis
- Enable synthetic metagenomics where thousands of genes identified in metagenomic studies can be "functionalized" through DNA synthesis followed by transfer and expression in appropriate hosts
- Organize, coordinate, and facilitate teams of hundreds to thousands of investigators to carry out large scale studies exploiting the JGI's technical, analysis and project management capabilities
- Provide a place where naïve users can come to work with experts dedicated to helping them solve complex energy and environmental problems requiring a combination of the wide spectrum of sophisticated JGI experimental and analysis capabilities
- Enable genome construction where bacterial genomes can be designed and synthesized for testing hypotheses
- Create a setting where an individual studying a specific organism will be able to annotate every gene in the organism's genome by accessing the JGI's portfolio of whole genome functional analysis capabilities
- The ability to carry out "extreme" metagenomics investigations where petabases of data are generated and then interrogated using sophisticated algorithms running on high performance computing infrastructure

A DRAFT STRATEGIC VISION

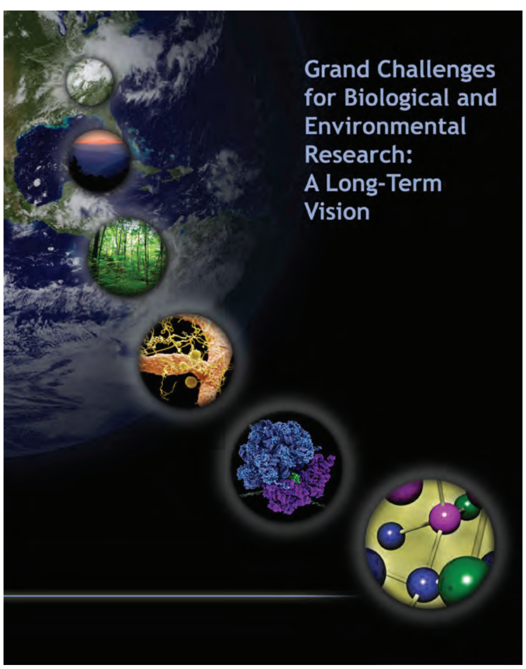
The present document is labeled as a "draft" 10-year strategic vision and represents a completed preliminary outline of the DOE JGI's future. It is, however, based on a substantial amount of input from multiple panels of external scientists as well as a broad representation of the DOE JGI user community (see Appendices). This "draft" plan will thus provide guidance for near-term decision making at the DOE JGI and is expected to serve as a mature and robust foundation for the development of a refined 10-year strategic vision following review and additional input from external communities of scientists.

II. BACKGROUND – SCIENCE DRIVERS

The primary focus of this strategic plan is to envision the capabilities of a next-generation genome center. These capabilities are driven and defined by the central scientific questions of energy and environmental research. In this section, we highlight the major areas that we believe will represent the science drivers for the DOE JGI over the next decade.

MISSION AREAS

As the DOE JGI continues to evolve, it remains deeply committed to the central missions of the Office of Biological and Environmental Research (BER) in biofuels, global carbon cycling and management, and stewardship of contaminated DOE sites. The BER Advisory Committee (BERAC) recently produced a strategic planning document, *Grand Challenges for Biological and Environmental Research: A Long Term Vision.* This document presents high-level views of BER goals in energy security, climate change and environmental remediation, as well as a specific discussion of strategic capabilities in systems science, computations, and science education relevant to BER missions. The DOE JGI's strategic direction outlined in this planning document dovetails with BER scientific goals and research areas.



BER Advisory Committee (BERAC) strategic planning document, "Grand Challenges for Biological and Environmental Research: A Long Term Vision."

ENERGY SECURITY

SCIENCE DRIVERS FOR THE DOE JGI NEXT-GENERATION GENOME CENTER – ENERGY SECURITY

- Double, over 20 years, the share of energy needs met by bioenergy in environmentally and economically sustainable ways
- Develop affordable and competitive options for energy supply and conversion that minimize negative impacts on climatic, environmental and ecological systems

Grand Challenges and Research Recommendations identified by the BERAC 2010 Long-Term Vision

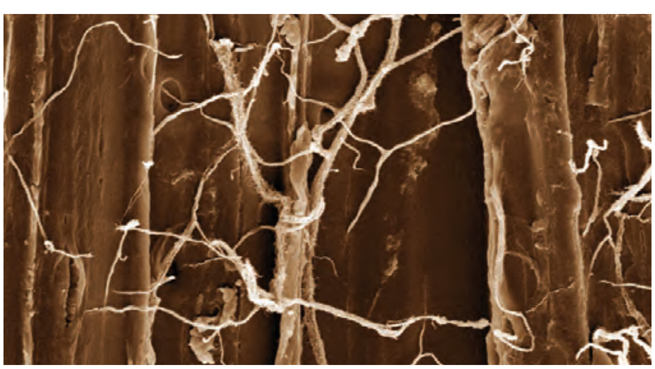
The BER long-term vision is founded on the realization that maintaining energy security and standard of living in the United States over the coming decades requires significant progress in energy efficiency and in developing energy supplies that are independent of imported fossil fuels. Transportation fuels are the highest-level priority BER targets because they are currently derived almost exclusively from petroleum. A well-defined strategy to produce cellulosic biofuels from biomass has been elaborated (*2011 U.S. Billion-Ton Update: Biomass Supply for a Bioenergy and Bioproducts Industry*) and the first steps have been taken including the funding of three DOE Bioenergy Research Centers. The BER long-term vision proposes to "triple the amount of degraded land on which perennial energy crops are planted in a manner that increases soil carbon storage and water quality" in order to double the fraction of total energy needs met by bioenergy over the next 20 years.

If these goals are to be met, biomass crops must be substantially improved to increase yields while growing on marginal land with minimal inputs of water and fertilizer. Historically, crop improvement has been slow and inefficient, requiring numerous rounds of breeding and selection of mature plants to produce a crop plant with desired traits.

Soybean Field (Credit: USDA).

The advent of high-throughput genome sequencing and sophisticated statistical genetic strategies now allow this process to be dramatically shortened through identification and selection of desired traits based on high-throughput genotyping. In addition, thorough understanding of the systems biology of plant-microbe interactions will enable maximization of growth, minimization of water and fertilizer use, and reduction in disease susceptibility. Genomic technologies that enable systems-level understanding of biomass feedstocks and their associated microbes will continue to be an area of emphasis for the DOE JGI.

Biomass deconstruction and extraction of fermentable sugars remains a critical bottleneck in the cellulosic biofuels strategy. Currently less than 50% of sugars contained in plant cell walls are extracted by industrial processes which restrains progress that can be achieved through development of high-yield cellulosic biomass crops. Recent studies have begun to elucidate the variety of mechanisms employed by microbes in nature to degrade biomass and there is little doubt that novel pathways and activities found in nature can be discovered and industrialized. The ongoing genomic revolution facilitates deep environmental sequencing, assembly and annotation of novel organisms and high-throughput testing and modification of genes and pathways. The tools of this revolution—gene and microbial engineering—will be central capabilities of the DOE JGI.



Postia placenta (Credit: Tom Kuster, FPL).

Finally, if biofuels are to replace a significant fraction of petroleum-derived fuels, they must be produced cost-effectively *and* be compatible with the current transportation fleet. Thus, an important target will be the microbial production of long-chain hydrocarbons from plant cell wall-derived sugars. Organisms discovered in the wild will not meet this target and significant modifications will need to be made to their genomes to increase carbon flow into desired products. Here again, gene discovery from environmental organisms, metabolic pathway analysis and subsequent manipulation using the tools of gene and microbial engineering developed at the DOE JGI will be absolutely essential to the success of developing better biofuels.

ENVIRONMENT AND CLIMATE SOLUTIONS

A decade into the 21st century, a recurrent theme in the scientific and mainstream discussion is the realization that anthropogenic climate change represents a serious and imminent threat with potentially catastrophic consequences for humankind. A large and growing volume of evidence underscores the need for immediate action. Emphasis has now shifted from assessing whether the risk is real, to applying cutting-edge science to develop creative and effective solutions to better understand and mitigate the threat. The essential role of plants and microbial communities in modulating climate change is widely accepted, yet most of the details of how these processes occur are not understood. In the future, the DOE JGI will focus on deploying experimental and computational technologies required to accelerate our understanding of these natural processes. Exploration of the systems-based interactions of plants and microbes in experimental and natural environments will contribute to an understanding of basic processes that will determine the trajectory of climate and environment in our changing world.

Over the next two decades, computational models of Earth systems are expected to dramatically improve, allowing modeling of terrestrial environments at unprecedented resolution. For such models to be meaningful and accurate, it will be important to understand carbon, nutrient and toxin fluxes in the Earth's environments (including oceans, permafrost, soils, cloud aerosols, terrestrial and oceanic vegetation) in much greater detail. If effective descriptions and predictions are expected, the ability to process, characterize and compare large numbers of environmental samples— key capabilities of the DOE JGI—must be tightly coupled with other experimental data derived from these environments. The DOE JGI, in positioning itself at the interface of genome analysis and experimental biology, can contribute to the integration of these increasingly large datasets in order to make them accessible and maximally useful to climate scientists and the modeling community.

Microbes occupy a central role in the cycling of carbon and nitrogen within terrestrial ecosystems. Their presence is responsible for the decomposition of organic matter and mineralization of nitrogen, and for additional interactions that arise between plants and climate that are mediated by the physical and chemical properties of soils. These cross-disciplinary connections increasingly drive insight into climate change. Thus, plants, microbes, and biogeochemical cycles are central to understanding the fate and transformation of carbon and nitrogen in soils of globally important, sensitive, yet poorly understood natural systems. In soils, characterized by large carbon and nitrogen stocks, microbial community composition may dictate whether tropical and boreal forests, peat lands, or high-latitude Arctic tundra will be net sources or sinks of greenhouse gases in the coming decades and whether these ecosystems will have large-scale positive or negative feedbacks to the state of the climate.

How best to quantify and represent these processes in models is uncertain and will require an ambitious research agenda linking climate science, biogeochemical cycles, and knowledge not only of structural, but also functional activities encoded in the genomes of plant and microbial communities, as well as mechanistic details of how plants and microbes interact. This is a formidable challenge, but one that can be met by emerging systems-biology research for which the DOE JGI and its varied genomic technologies will need to serve as a central hub.

SCIENCE DRIVERS FOR THE DOE JGI NEXT GENERATION GENOME CENTER – CLIMATE AND ENVIRONMENT

- Develop higher-resolution models in order to integrate many more relevant processes than offered by current models and to describe climate change over much longer time scales
- Develop ecosystem-observing systems to monitor biogeochemical cycles, estimate critical process parameters, and provide model tests in ocean and terrestrial biospheres, including subsurface soils
- Advance understanding of important biological interactions and feedbacks to identify potential tipping points and possible mitigation strategies such as carbon biosequestration

Grand Challenges and Research Recommendations identified by the BERAC 2010 Long-Term Vision

Stone Lakes National Wildlife Refuge in California (Credit: Justine Belson/USFWS via Flickr).

STRATEGIC CAPABILITIES

TOOLS FOR SYSTEMS SCIENCE

At the core of BER's long-term strategy outlined in the previous pages is a higher-level understanding of individual organisms and their interactions between themselves and their environments. With this understanding, more complete predictive models can be constructed and tested, and activities, pathways and organisms can be engineered to solve important problems in energy and the environment. This higher-level functional understanding is also the explicitly stated goal of modern systems science. Developing and exploiting the tools of systems biology for functional understanding is front and center in the BER long-term vision and the DOE JGI's strategic plan.

PREDICTING PHENOTYPE FROM GENOTYPE

The key to understanding the molecular basis of robustness, fitness and selection is the ability to characterize and link inter-species or population genetic variation to specific phenotypes. For example, comparative strategies have been successfully applied to identify single genes essential for long-chain alkane synthesis in cyanobacteria and for associating multiple genes with complex traits like grain yield in rice. However, to meet the ambitious goals set forth in the long-term BER vision, genome comparisons must be extended to encompass genetic variation existing in environmental microbial consortia and epigenetic variation among single cells isolated from complex organisms. Similarly, phenotyping efforts must be refined and expanded to capture the organizational rules that govern interactions between organisms in the plant rhizosphere, biomass-degrading environments, and other marine and terrestrial environments. Finally, the statistical genetic tools for determining genes responsible for specific traits in plants, microbial isolates and microbial communities must be refined and made widely available to the DOE user communities.

SCIENCE DRIVERS FOR THE DOE JGI NEXT-GENERATION GENOME CENTER – SYSTEMS AND SYNTHETIC BIOLOGY

- Determine the molecular basis of robustness, fitness, and selection
- Apply advanced computational and analytical capabilities to characterize the information molecules and network interactions used by biological systems
- Understand, predict, and manipulate the types and rates of ecosystem responses that influence climate change
- Deploy synthetic biology (biodesign) to understand and manipulate ecosystem function

Grand Challenges and Research Recommendations identified by the BERAC 2010 Long-Term Vision

ANALYTICAL AND COMPUTATIONAL CAPABILITIES TO CHARACTERIZE GENE NETWORKS AND METABOLIC PATHWAYS

Genes operate in regulatory networks to fine-tune organism growth, development, response to stress and cell differentiation. Nevertheless, even in relatively simple microbial organisms our current ability to provide complete functional annotation of individual genes is limited. Placing genes into metabolic pathways and regulatory networks provides a higher level of annotation than simple comparison or activity assays can provide. This will be essential if we are to be able to predict an organism's phenotype and its responses to environmental influences. Similarly, while many of the fundamental principles of integrated gene network analysis have been established, better analytical and computational methods are needed for higher-resolution spatial and temporal characterization of gene expression, gene regulatory networks, and dynamics of regulatory responses to diverse stimuli. The DOE JGI will develop and apply both sequence-based and experimental approaches to define gene regulatory networks and their dynamics at high spatial and temporal resolution.

EXTEND PREDICTIONS FROM SIMPLE MODEL SYSTEMS TO MORE COMPLEX "REAL-WORLD" ENVIRONMENTS

Establishing model systems with sequence- and phenotype-derived deep functional genome annotations will allow predictions of how organisms respond to change. When combined with an understanding of population and community structures, these data sets will also enable a deeper understanding of the evolution of these organisms. These models can be extended to include interactions with other organisms as in low-complexity microbial communities, and gradually to more complex systems. At each stage, it is imperative that predictions from model systems be compared to observations from natural habitats and environments to enable iterative improvement in annotations and predictions of responses. The DOE JGI has a long-standing interest and expertise in the functional annotation of genomes. The ability to capture and present phenotype information in genome annotations for access and utilization by the user community will thus be essential for continued success in this area.

DEPLOY GENE SYNTHESIS TO UNDERSTAND AND MANIPULATE ECOSYSTEM FUNCTION

The dramatic progress in our ability to read DNA sequence and the availability of massive amounts of digital sequence information in public repositories has led to a growing interest in the ability to synthesize DNA on the scale of single genes and complete pathways. In its simplest form, gene synthesis allows high-throughput expression and testing of gene function in the absence of clones derived from the sequenced organisms. Gene synthesis also allows sequences to be modified to alter expression or activity. However, the game-changing potential of high-throughput DNA synthesis is the ability to design large DNA fragments spanning tens or hundreds of kilobasepairs that will allow entire metabolic pathways to be transplanted from one organism to another, that include substantial changes to optimize and control expression of each component to avoid toxicity and maximize output of desired product(s).

Island Lake, Alberta, Canada (Credit: Gord McKenna/Flickr).

Currently, there is considerable potential for improvement of synthesis methods. If technical obstacles can be overcome, synthetic strategies offer a conceptually straightforward and technically realistic approach to harness the predictive power of next-generation genome analysis to control the flow of carbon into biomass for biofuels or sequestration applications, to enhance biomass degradation in biofuel production, to synthesize fuels and other industrially important intermediates, or to metabolize environmental toxins. The DOE JGI next-generation genome center outlined in this document will make significant strategic investments in gene synthesis and biodesign.

COMPUTING FOR BIOLOGICAL AND ENVIRONMENTAL RESEARCH

SCIENCE DRIVERS FOR THE DOE JGI NEXT-GENERATION GENOME CENTER – COMPUTING AND ENVIRONMENTAL RESEARCH

- Develop new computing paradigms capable of meeting the enormous parallel processing and data-intensive analysis needs now emerging for biological, climate, and environmental data
- Standardize experimental and computational protocols and methods to increase data integration, data usability, and system interoperability to improve research productivity
- Design and build software solutions that provide researchers with better access to increasingly large, complex, and interrelated datasets

Grand Challenges and Research Recommendations identified by the BERAC 2010 Long-Term Vision

NEW COMPUTING STRATEGIES FOR DATA-INTENSIVE SCIENCE

As sequence production continues to expand at exponential rates, algorithms for assembly, annotation and large-scale comparison must be developed on the assumption that they will be ported to high performance computing environments. Currently, many essential algorithms like those used for assembly are both memory intensive and difficult to parallelize. Innovative programming strategies and potentially redesign of these algorithms will be required to utilize these high performance computing environments. For effective presentation to users these large datasets will need to be compressed, while preserving the complexity of the underlying systems.

STANDARDIZED COMPUTATIONAL PROTOCOLS

As sequencing and analysis tools become commodities, the fraction of the world's data that the DOE JGI produces will shrink. The DOE JGI must actively engage other data producers to develop standards that allow data generated by the DOE JGI and the community to be integrated, exported to and used by the DOE's Knowledgebase and other biological analysis systems that will inevitably arise. This means that production of the primary data, as well as the algorithms for assembly, annotation, variation detection and gene expression, must be standardized and routinely benchmarked, with caveats and possible errors flagged for users through careful documentation.

SOFTWARE SOLUTIONS FOR LARGE, COMPLEX, AND INTERRELATED DATASETS

Development of automated tools for genome assembly, annotation, analysis of gene expression and proteomic applications is proceeding today at a rapid rate. However, there is much more to do if systems science at a large scale is to be enabled. Challenges that remain include: deciphering the function of the majority of genes we discover, predicting how genome variation will affect gene expression or function, reliably collating genes into functional pathways and regulatory networks and understanding how organisms interact within even the simplest systems.

The problem of providing these functional descriptions is ultimately tractable, but the solution will require acquisition of very large experimental datasets and the development of robust analytical software tools that allow this new information to be served to users in a meaningful way. High-throughput functional screens will lead to reliable computational prediction of gene function. Standardized methods for expression analysis, proteomics and metabolomics must be translated to a standardized representation of metabolic pathways and regulatory networks in plants and microbes. Statistical genetic algorithms developed for humans must be implemented and refined to associate plant genome variation with phenotypes. Finally, experimental and computational methods to analyze how organisms interact with each other must be developed if functional predictions about individual environments or the biosphere are to be attempted.

The DOE JGI has played a leadership role in developing, standardizing and providing access for users to high-quality genome assemblies, annotations and comparative genomics tools. The DOE JGI is uniquely positioned through its interaction with users to expand this repertoire in pursuit of systems-level analysis on behalf of its users.

Finally, at present there are relatively few computational biologists who are trained to operate in high-performance computing environments and even fewer experts in high performance computing who understand the challenges faced by biologists. It is essential that the DOE JGI and DOE's supercomputing facilities make the bridging of this gap a high priority and allocate resources to it.

II. CAPABILITIES

In response to the Science Drivers outlined in the Background section, the primary focus of the DOE JGI's vision for the future will be the active development and application of state-of-the-art genomic capabilities to enable rapid progress in these areas.

In this section, we have divided these capabilities into three general categories or "pillars":

Pillar 1: Experimental Data Generation

Pillar 2: Biological Data Interpretation

Pillar 3: User Interactions

For each pillar, we define a set of goals and stretch goals, describe the strategies required to accomplish them and provide examples of applications for these new capabilities.

As a user facility, the DOE JGI will put continued emphasis on enabling its users to harness these cutting-edge capabilities to solve the critical DOE mission-relevant science problems of the next decade.

PILLAR 1: EXPERIMENTAL DATA GENERATION

EXPERIMENTAL DATA GENERATION – STRETCH GOALS

- Routinely assign a function to >90% of the genes in microbial and plant genomes
- Discover new branches of life through metagenomics and massive-scale single-cell sequencing of uncultured organisms
- Characterize and model complex environmental systems to a level where we can correctly predict response to changes
- Design and build a genome to address a whole-genome hypothesis

EXPERIMENTAL DATA GENERATION – GOALS

- Sequencing: Leading-edge molecular biology and sequencing capabilities at multi-petabase scale
- Massive Scale Sample Preparation: Development of technologies for designing and carrying out complex "grand challenge" projects involving hundreds of collaborators, thousands of conditions, and tens of thousands of samples
- Single-Cell Analysis: Able to interrogate at a genomic, transcriptomic, proteomic and metabolomic level thousands of isolated single cells from plant tissues and mixed environmental samples
- Single Chromosomal Genomics: Able to separate, isolate and interrogate single plant chromosomes
- Synthesis: Able to synthesize and express thousands of genes and large pieces of DNA to engineer complex organisms for hypothesis testing
- Functional Annotation of Microbial Genomes: Genome-wide perturbation and read-out of microbes under multiple conditions

EXPERIMENTAL DATA GENERATION – SCIENCE STRATEGY

- Expand state-of-the-art sequencing capabilities
- Develop automated sample processing and tracking that captures all steps from field collection of meta-data to sequencing to database storage
- Develop automated single-cell sorting and processing
- Develop high-throughput single chromosome capture and analysis
- Develop robust automated genome amplification and RNA processing strategies
- Develop high-throughput DNA synthesis capabilities to build large pieces of DNA
- Develop whole-genome synthesis approaches
- Develop high-throughput microbial genome saturation mutagenesis and phenotyping

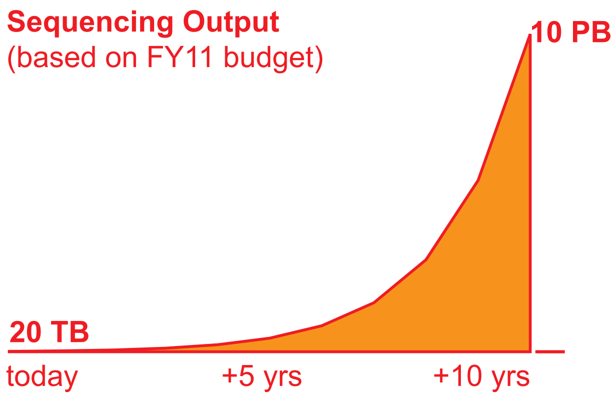
KEY:

Advance existing capabilities
Expand nascent capabilities
Develop new capabilities

A. SEQUENCING

OVERVIEW

The DOE JGI was formed in connection to the Human Genome Project for the primary purpose of providing cost-effective high throughput DNA sequencing that could only be accomplished by scaling of production activities at a centralized facility (1998-2004). In its second incarnation as a user facility (2005-2011), major accomplishments resulted from the partnership of outside scientists leveraging genomic expertise available centrally through the DOE JGI.



Sequencing will continue to be a core capability of the DOE JGI. Owing to dramatic decreases in sequencing cost, the budgetary limitations of large-scale projects will increasingly be defined by the cost of computation and downstream analysis. Nevertheless, due to continuous and predictable cost improvements of computational analysis, the sequence output achievable with a given budget is expected to increase by approximately three orders of magnitude over the next decade.

A key factor in the success of the DOE JGI was not only having state-of-the-art DNA sequencing and analysis capabilities, but also having cutting-edge scientific staff to aid, (and in many cases lead) both upstream sample processing and downstream sequence analysis. These areas have proven to be critical to the success of user projects and both are expected to continue to grow in the future.

We anticipate DNA sequencing technologies to continue to generate more data per unit cost and linked to these capabilities, we plan to integrate upstream sample processing workflows, as well as downstream analytic capabilities. Areas upstream of sequencing will be focused on scaling DNA isolation and library construction through continued improvements in automation and single-cell molecular biology to enable studies of environmental samples and complex organisms at unprecedented granularity. Sequencing will also be tightly linked to the emerging synthesis capabilities at the DOE JGI, to enable and support the creation of large synthetic DNA constructs.

CONTINUED LEADERSHIP IN SEQUENCING APPLICATIONS

As the DOE JGI transitions from a production sequencing facility to a next-generation genome center, a major thrust in this strategic plan is to add new genomic capabilities to the DOE JGI. However, all of these efforts are founded on the continued top priority of maintaining and expanding the DOE JGI's core world-class leadership in nucleic acid sequencing. It is a critical strategic target that the DOE JGI has massive-scale sequencing capabilities available to accommodate the anticipated substantial improvement in sample throughput and user demand for sequence generation.

B. SAMPLE PREPARATION

OVERVIEW

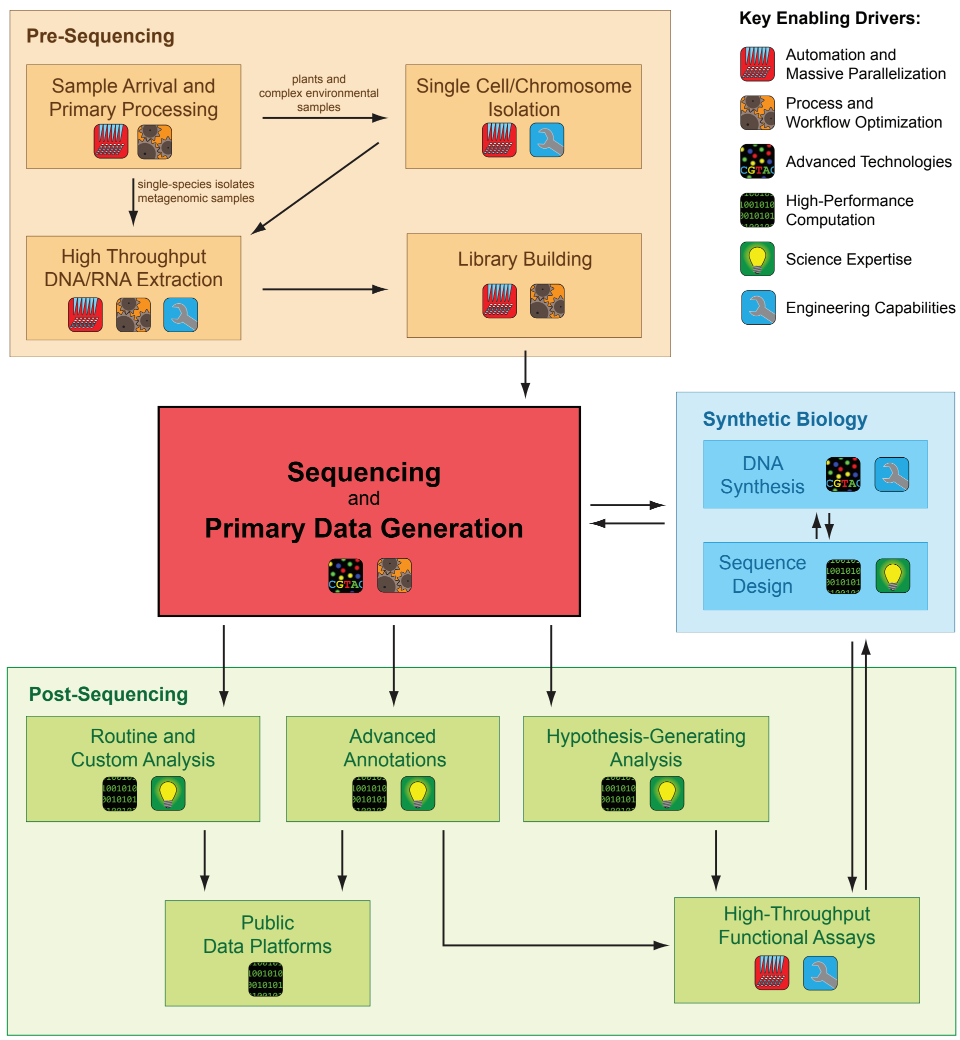
While the output of nucleic acid sequencing has increased by orders of magnitude, obtaining adequate DNA and RNA samples in sufficient numbers continues to be a challenge for the DOE JGI and its user base. This is frequently due to limitations in sample quantity or quality, but also results from the inability to prepare DNA or RNA in a high-throughput manner. To overcome this pressing limitation, one area of strategic focus is to provide routine access to a variety of automated sample preparation capabilities at the DOE JGI. Two examples from the growing list of such required capabilities that will be implemented over the coming years are large scale automated sample preparation and nanogram quantity sequence assays.

LARGE SCALE AUTOMATED DNA/RNA SAMPLE PREPARATION

To date, the DOE JGI has relied upon users to isolate DNA and RNA from starting materials. In the future, we anticipate adopting automated solutions for extracting nucleic acids from hundreds to thousands of homogeneous, heterogeneous and environmental samples. It is the inherent diversity of these samples that sets DOE JGI's effort apart from corresponding standardized sampling of human DNA/RNA for biomedical applications. DOE JGI users will require large panels of material from plants or soil for DNA surveys of genetic variation or metagenomics, respectively. Similar pursuits in RNA extraction will be developed to aid in exploring large numbers of tissues or environments for gene expression counting. These pursuits are in line with core DOE JGI expertise and will require informatics support, extending sample management, liquid handling and robotic automation/engineering, as well as novel molecular biology development for DNA/RNA extraction from a variety of biological materials. It is anticipated that such capabilities will allow studies involving massive numbers of samples to be carried out at the DOE JGI by standardizing and centralizing DNA and RNA extractions at scale. For example, substantial opportunities exist in the resequencing of thousands of individual trees from a forest to relate variation to phenotype and to massively expand our knowledge of environments through large-scale metagenomic sampling. Each of these types of project will require the careful coordination of sample processing and management, high quality sequence generation, as well as integrated sequence analysis.

NANO-GENOMICS

Another front-end capability that will expand our access to more readily available samples for nucleic acid sequencing is to continue to push the limits on the minimal amount of starting material required for existing protocols at the DOE JGI. Presently, the vast majority of methods require micrograms of starting material while many users remain challenged to even obtain nanograms from certain biological environments. Solutions are expected in microfluidics and emulsion oil-based miniaturization to provide an entry point for many more project types, without the large biases seen with standard amplification methods. Based on the DOE JGI's expertise in molecular biology and technology development, these are well-aligned areas for DOE to continue to provide scientific leadership.



Sequencing as a core capability at the DOE JGI is tightly linked with other advanced capabilities of a next-generation genome center.

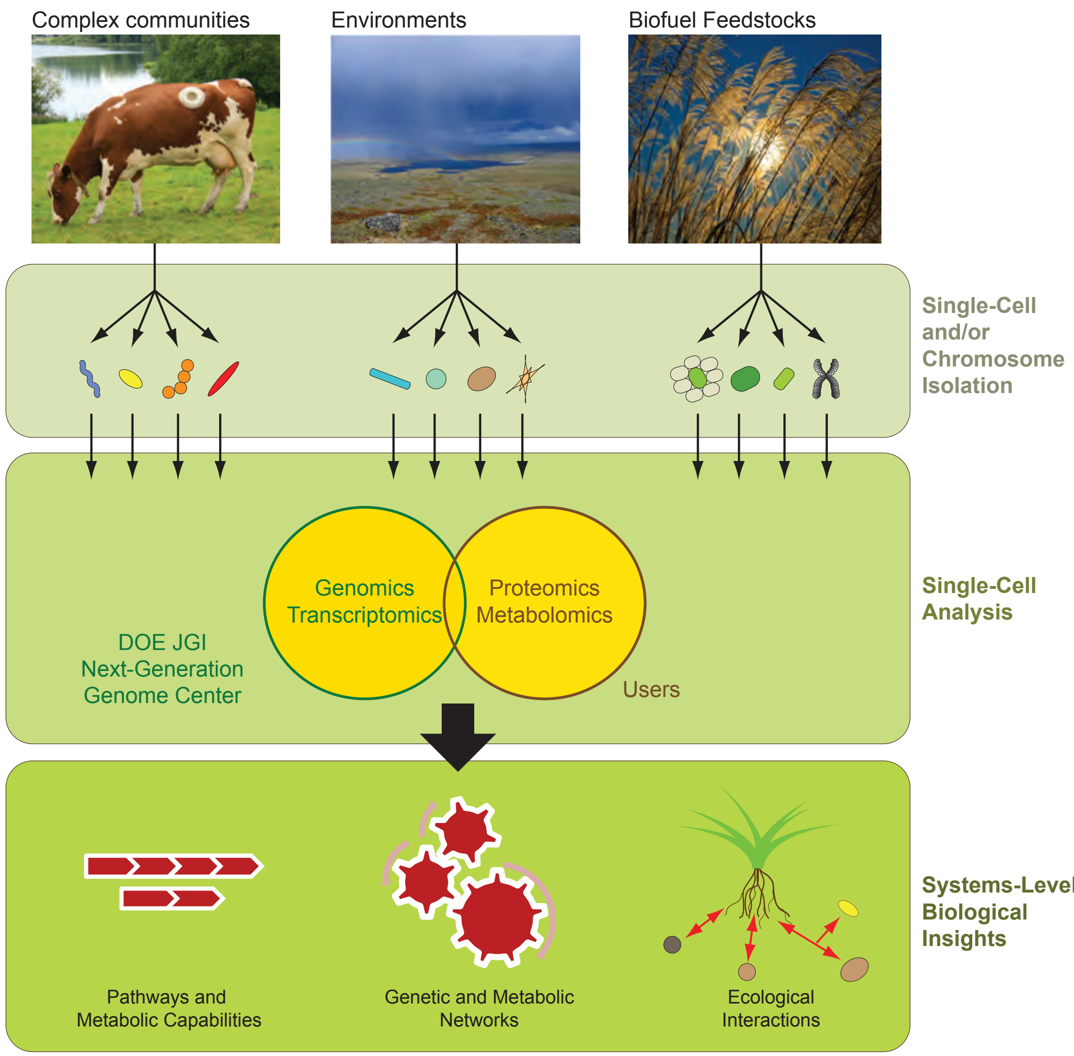
C. SINGLE-CELL AND SINGLE-CHROMOSOME GENOMIC ANALYSIS

OVERVIEW

Only a minute fraction of microbes can currently be cultured *in vitro*, representing a substantial obstacle for exploring the biology of the vast majority of microbes. Importantly, this includes large numbers of microbes that are relevant to energy and environmental applications.

Culture-independent approaches such as metagenomics and, more recently, metatranscriptomics have provided a first path into an understanding of the uncultured microbial biosphere, tackling many questions of DOE relevance. However, most of these techniques have considerable limitations for exploring individual species that are members of heterogeneous and often complex ecological communities.

Emerging single-cell technologies provide a powerful complementary strategy to access the genetic make-up of individual uncultured community members, eliminating key challenges of metagenomic approaches, such as the proper assembly and binning of complex data sets.



Single cell isolation capabilities paired with highly sensitive and high throughput 'omics' approaches will enable a transition towards systems biology at the single-cell resolution, providing an unprecedented view into environmental cellular pathways and networks.

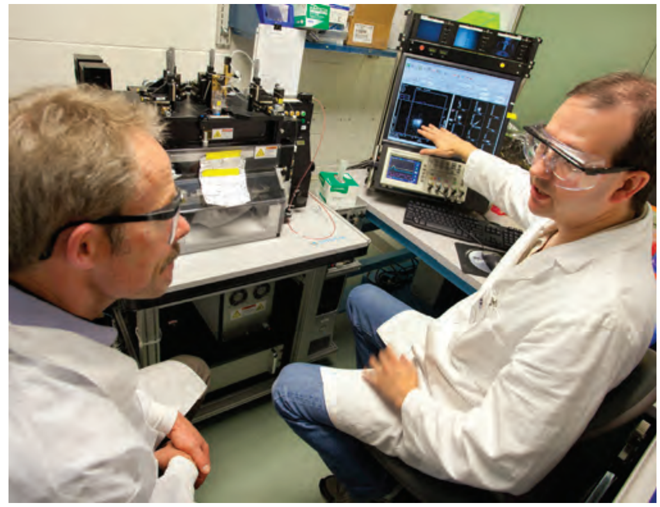
Current single-cell technologies allow the recovery of genomes and transcripts from uncultured individual prokaryotic and eukaryotic cells, providing a link between phylogeny, metabolic potential and expression activity. Moreover, single-cell metabolic, proteomic and peptidomic analyses have recently been accomplished in eukaryotes, providing deeper insights into cellular networks and circuits. While these methods are currently technically challenging, tedious and low in throughput, they provide the first step towards a single-cell level systems biology understanding of life on Earth.

At the DOE JGI next-generation genome center, large-scale single-cell analysis will be established as a key strategic capability to enable users to leverage the potential of these techniques for energy and environmental studies. Further method development will be aggressively pursued to mitigate the technical challenges that currently still limit the throughput of single-cell techniques. In particular, methods will be streamlined using micro- and nanofluidics approaches to increase sample throughput by orders of magnitude. This will also enable cell preparations for complementary single-cell proteomics and metabolomics studies of the same specimens by users, in order to enable systems-level studies at single-cell resolution.

The highly polyploid or polymorphic eukaryotic genomes invariably associated with bioenergy-relevant feedstocks pose major challenges for *de novo* whole genome sequencing projects. The sequencing of single chromosomes offers a means to simplify the assembly and make such undertakings more tractable. Moreover, high throughput isolation and sequencing of single plant chromosomes could enable large-scale haplotype studies. The access to single-chromosome genomic capabilities will immediately alleviate some of the current obstacles in the interrogation of plant genomes and would likely also enable applications in fungal and environmental genomics.

1. SINGLE-CELL GENOMICS

The lack of the ability to culture key players in highly critical environmental processes, such as carbon cycling and bioremediation, emphasizes the urgency of culture-independent methodologies to study these organisms on the nucleic acid, protein, and metabolite level. Performing this on the single cell level provides simplified datasets that will allow unprecedented insights.



User interactions (Credit: Roy Kaltschmidt, LBNL).

Efforts to establish large-scale single cell technologies at the DOE JGI are based on a firm conceptual, scientific and technological foundation. Over the past few years, the DOE JGI has developed first moderate-scale single cell genomics pipelines. Great strides have been made in providing proof of concept that single cells can be isolated, amplified, screened, sequenced, and their genomes assembled and analyzed at our facility, yielding high quality assemblies or even complete genomes from single cells of uncultured microbes.

A major strategic target for the next decade is increasing automation and streamlining of all steps in single-cell genomics pipelines, with the goal to be able to handle tens of thousands to millions of single cells in a day, adapt the pipelines to different sample types (including plant tissues and single eukaryotic chromosomes) and expand them to complementary systems-level approaches including preparation of single cells for metabolic or proteomic studies. The DOE JGI is well positioned to accomplish these goals with infrastructure and expertise in single-cell handling and molecular methods dealing with extremely small template amounts. Investments are expected in miniaturization of sample handling to move beyond micro-titer plates and into higher throughput workflows, as well as in the implementation of targeted single cell approaches. By being able to isolate single cells, nuclei or chromosomes of interest on a production-scale, we expect to be able to reduce the complexity of metagenome samples to aid in simplifying analytic methodologies and to obtain microorganisms that are biologically important, but difficult to obtain. This coupling of high-throughput single cell capabilities with transcriptomics, proteomics and metabolomics is expected to yield substantially improved insights and comprehensive systems biology level views of our environment.

UNIVERSAL HIGH THROUGHPUT METHODS TO REDUCE SAMPLE COMPLEXITY

A major capability request from DOE JGI users is in the area of selective cell collection from complex biological samples, such as specific cells types from plant tissues or the isolation of microbial species associated with plant root systems (the "rhizosphere"). One of the DOE JGI's key roles will be to provide the users with a portfolio of technologies including cell sorting, micromanipulation, and laser microdissection, enabling the isolation of their community samples, tissues, cells, nuclei or chromosomes of interest, followed by their downstream processing and 'omic' analysis.

MINIATURIZATION AND STREAMLINING

The goal of the DOE JGI will be to implement and maintain state-of-the-art technologies to adopt single-cell and single-chromosome sequencing into miniaturized workflows. As "next next-generation" sequencing technologies are expected to evolve, a future vision is the sequencing of genomes and/or chromosomes of individual cells without the need for pre-amplification and/or library generation, opening the door to sequence analysis of native DNA within individual cells. The DOE JGI will be at the forefront of such novel endeavors decoding native genomes of individual cells within hours, deciphering their genetic code and potentially unraveling structural modifications of hitherto unknown biological significance that become evident at single cell resolution.

TARGETED SINGLE-CELL APPROACHES

In current high throughput single cell genomic experiments, cells are randomly chosen for whole genome amplification, followed by identification using 16S rDNA PCR screening. While ribosomal RNA-targeted fluorescence in situ hybridization (FISH) can be used to isolate particular phylotypes of interest on the single cell level, high throughput pre-screening for cells encoding particular single or low copy number genes of interest has not yet been accomplished. A long-term strategic goal will be the development of such targeted and universally applicable single cell technologies.

2. SINGLE CHROMOSOME GENOMICS

The majority of productive crop plants in the U.S. are either polyploids, whose genomes consist of more than two copies of each chromosome (corn, wheat, barley, alfalfa, cotton), hybrids of two similar organisms (sugarcane, tree crops) or retained paleopolyploids (soybean). One direct way of accelerating work on these species is to bin specific chromosomes using micro-dissection technology for each of the copies of every chromosome, amplify these single chromosomes and then collect dense sequence-based markers across each chromosome. Whole-genome shotgun sequencing strategies can then be applied to localize all of the resulting contigs to a specific copy of the polyploid genome. This will provide a direct resource for breeding and improvement in these biofuel species, as knowledge of the location of specific allelic and subgenome variation is needed for breeding programs to produce more efficient and pest tolerant varieties.

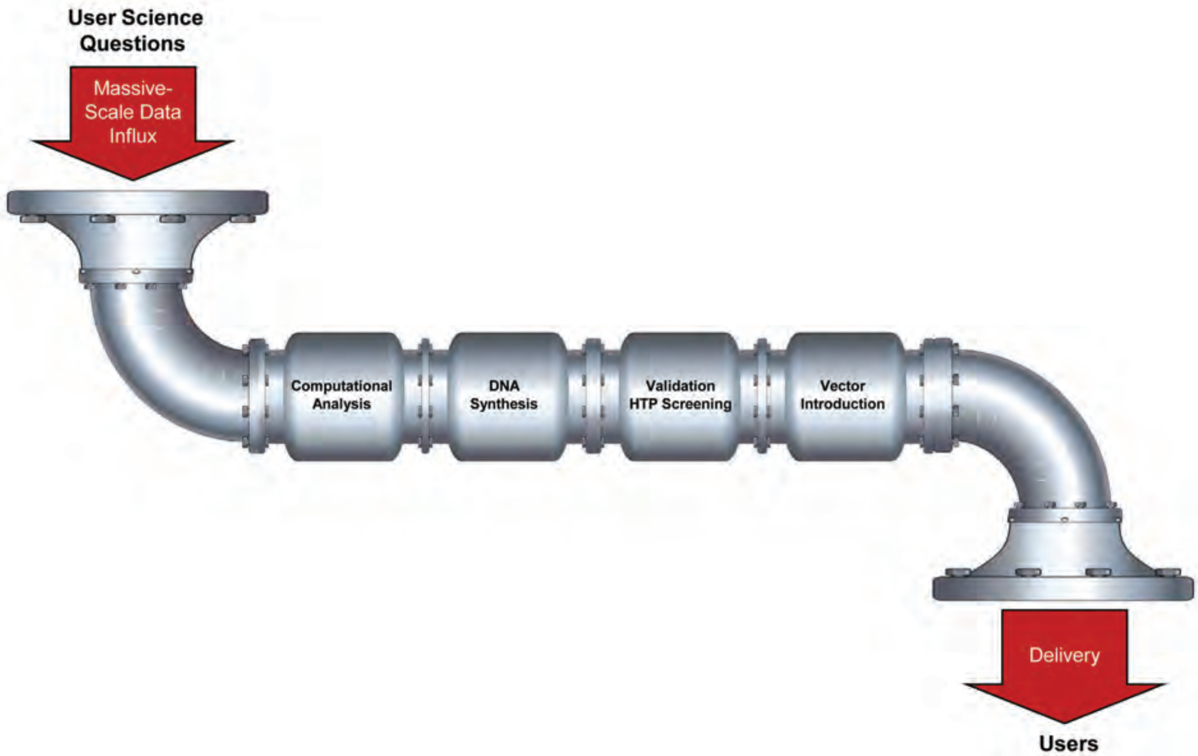
D. DNA SYNTHESIS FOR BUILDING GENES AND LARGE SEGMENTS OF DNA

OVERVIEW

Massive increases in DNA sequencing have provided the identity of millions of genes from environments of importance to DOE, with further massive increases expected in the future. However, deciphering the functions of these genes and other elements encoded in sequence data is lagging, particularly in the area of metagenomics. As an additional obstacle, deriving functional insight from metagenomic sequence studies is limited by the inability to recover the organism or physical DNA molecules represented by the digital sequence. Methodologies to easily obtain at a large scale these "digital" sequences as cloned DNA products that can be introduced into organisms for functional exploration will dramatically contribute to the deciphering of metagenomic data.

Advanced data mining of next-generation sequencing data combined with DNA synthesis and expression, complemented with downstream functional analyses, provides a powerful tool for large-scale characterization of 'digital genes.' The DOE JGI's vision is to develop an integrated pipeline from sequence data generation to user-driven functional characterization that leverages the institute's specialized capabilities in liquid handling, process optimization, automation, analysis, and sequencing which is a central capability in any DNA synthesis project. Most importantly, access to large scale DNA synthesis capabilities has been identified by DOE JGI users as a high priority for advancing their science in the future.

Sequence driven DNA synthesis solutions are currently out of reach for most investigators since they involve many complex steps including: sophisticated analysis of sequence data, designing gene and pathway constructs, building DNA fragments, and introducing DNA fragments into



DOE JGI integrated DNA synthesis pipeline for the conversion of digital sequence information into biological parts.

appropriate hosts that suit planned functional studies. The DOE JGI aims to improve synthetic methodology and make it available to scientists studying DOE-relevant problems, but lacking pathway and genome engineering capabilities. The ability to synthesize a large number of genes, as well as to combine them into larger DNA constructs will be a crucial capability for researchers in the future to link sequence to function in the study of environmental and energy-related problems. Based on the DOE JGI's long and positive interactions with users over the past 10 years and its technical and computational capabilities, it is well positioned to establish production-scale high-throughput design and assembly of DNA as a key capability offered to users to complement production-scale DNA sequencing.

SYNTHESIS CAPABILITIES TO BE ESTABLISHED AT THE DOE JGI

In order to enable users to perform sophisticated, large-scale synthesis projects, the DOE JGI's efforts in DNA synthesis will focus on two major areas:

1. Defining what to write: Data mining, construct optimization and design
2. *In silico* to *in vivo:* Producing and introducing the DNA into the organism of choice

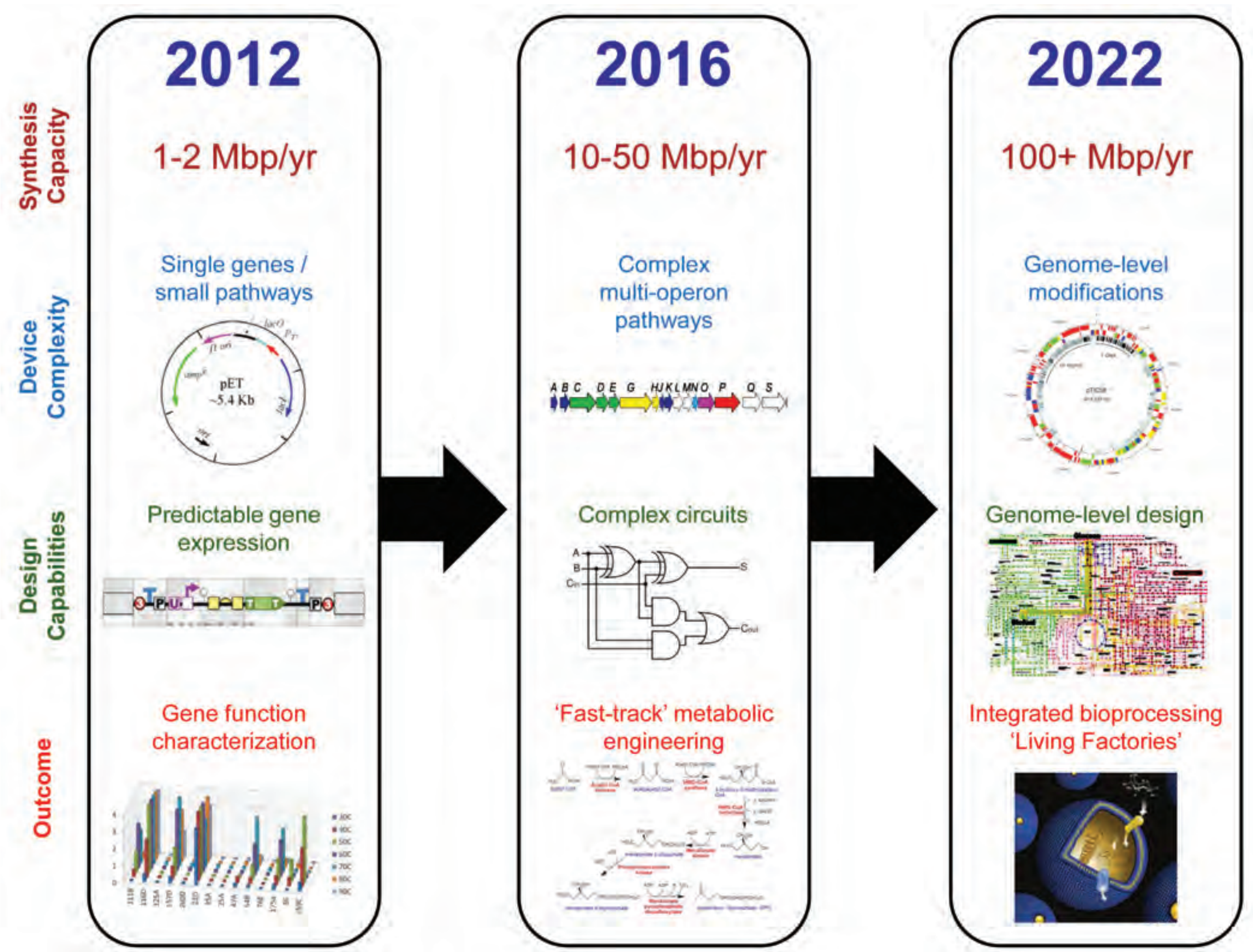
DEFINING WHAT TO WRITE: DATA MINING, CONSTRUCT OPTIMIZATION AND DESIGN

The first step in any synthetic DNA research program is to precisely define a set of sequences to be synthesized. This is itself a multi-step process involving: a) the mining of available sequence repositories for "raw" sequences with desired characteristics, such as encoding proteins with desired catalytic/structural properties or gene regulatory elements with desired response/activity profiles; b) the computational optimization of individual sequences, taking into account the properties of the eventual host system which may require optimization of codons, functionally neutral replacement of "prohibited" sequence motifs and features, and possibly hypothesis-driven alterations to change the function of protein-coding or regulatory sequences; c) devising a synthesis and assembly strategy that is compatible with the target sequence, as well as the characteristics of the assembly/host system.

IN SILICO TO *IN VIVO:* PRODUCING AND INTRODUCING THE DNA

Currently the most expensive and time-consuming step in large-scale synthetic DNA projects is the assembly of small oligonucleotides into larger fragments. The DOE JGI will be at the forefront of implementing new technologies into the DNA synthesis pipeline. These goals will be supported by the DOE JGI's longstanding experience in developing cutting-edge molecular processes, wet-lab automation and process optimization. The DOE JGI DNA Synthesis Pipeline will also benefit from the DOE JGI's state-of-the-art sequencing and computational analysis capabilities, which are central to this activity.

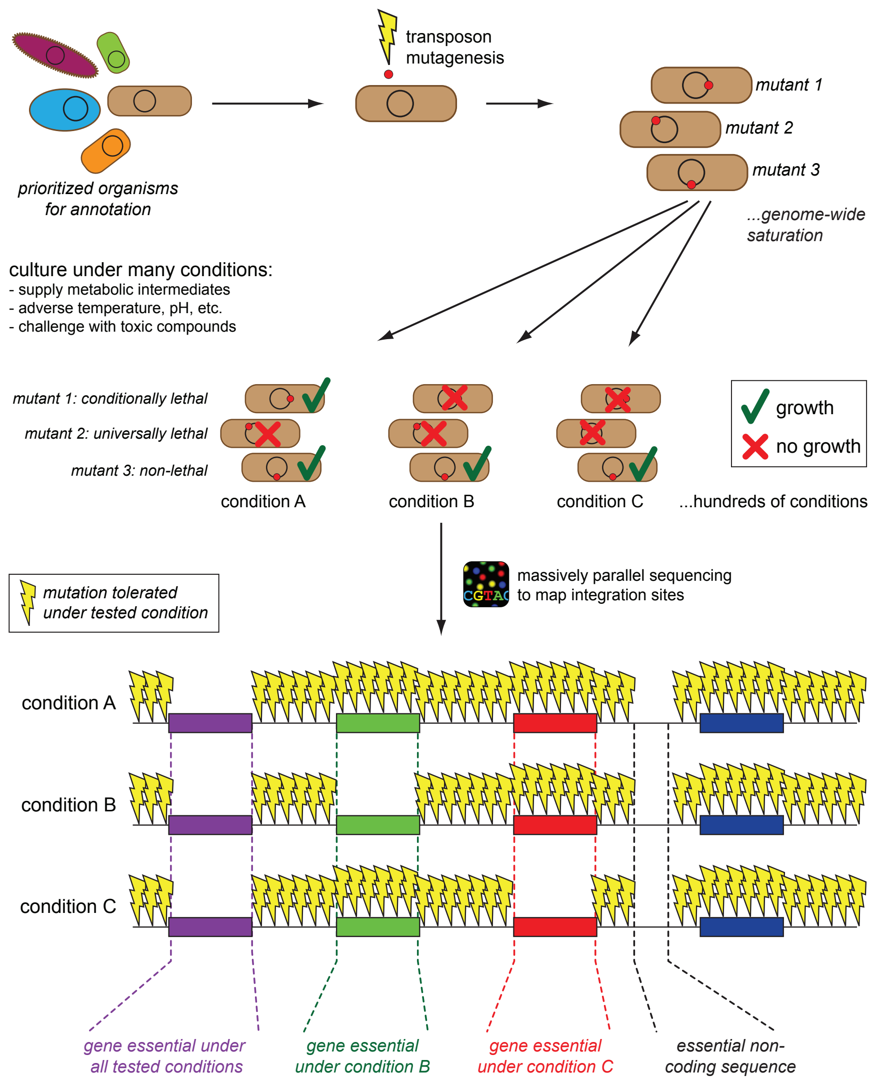
The DOE JGI's primary role in synthetic DNA projects will be to support users in the computational design of desired target constructs, in the creation of these large and complex DNA molecules and in their introduction into suitable host cells. In contrast, in-depth functional characterization of the resulting transformed host organisms will primarily rely on expertise and assays established in the respective users' laboratories. Nevertheless, in order to be able to support users in the ability to generate synthetic systems required to address energy and environmental challenges, the DOE JGI will also develop experimental paradigms in which functional readouts can be closely linked to synthetic sequence.



10 year vision for DNA Synthesis at the DOE JGI.

E. FUNCTIONAL GENOME ANNOTATION OF MICROBIAL GENOMES

Obtaining the sequence of individual microbial species is no longer the hurdle it once was. With technological progress, sequence generation and genome assembly of the data will cease to be a bottleneck in microbial and plant genome studies. In sharp contrast, complete functional annotation of genomes remains an unsolved challenge. This is poignantly epitomized by the current annotation status of the best-studied of all microbial species, *E.coli*. Even though *E.coli* and its genome have been studied for decades by countless laboratories, a significant fraction of its genes (~15%) still lack functional annotation. Obviously, this problem is far more severe for less well-studied microbial genomes, which make up the majority of microbes analyzed by the DOE JGI. Among these DOE-relevant species, typically 40-60% of all genes lack any annotation. The situation is worse for complex eukaryotes such as plants. As a result, a large proportion of the sequence generated is currently not useful for functional applications. Furthermore, inaccurate and incomplete annotations hamper systems-level understanding of organisms, including their metabolic properties and capabilities in the context of ecosystems related to energy and environmental challenges.



Pipeline for functional genome annotation by transposon-mediated saturation mutagenesis.

A variety of approaches have been employed for enhancing functional annotation of genomes including a myriad of computational and experimental approaches. Consensus has emerged that no single approach or environment will reveal the full repertoire of gene functions for a particular organism. Accordingly, the DOE JGI will develop a variety of different experimental and computational capabilities to enable complete and reliable whole-genome annotation.

Presently, only a small number of groups test genome-scale high-throughput experimental approaches (e.g. saturation mutagenesis) for adding functional annotation data to sequenced genomes. Due to the challenges of many of these technologies for most investigators, often only a single experimental approach is combined with an off-the-shelf computational annotation to examine their organism of choice. The DOE JGI next-generation genome center will offer a unique expertise in annotation and analysis of complex organisms and systems and a comprehensive suite of tools and techniques, both experimental and computational, to deliver near-complete functional annotation of sequenced genomes to our users. We will enable investigators to access not just a single annotation platform, but a suite of platforms, assisted by a skilled DOE JGI staff annotator. In these efforts, the DOE JGI will encourage scientists to partner with the DOE JGI to access whole-genome annotation capabilities to analyze their organism from this perspective. These studies will initially focus on intensively studied DOE-relevant microbes with the goal to assess what combination of approaches is most efficient and scalable. A stretch goal of these efforts is to produce microbial genomes in which nearly all genes have at least a basic level of annotation provided through a combination of comparative and experimental approaches.

ADVANCED CAPABILITIES FOR FUNCTIONAL ANNOTATION

The DOE JGI has already produced reference genomes for more than 500 bacteria/archaea, ~100 eukaryotic microbes and 15 plants. For many of them we utilized transcriptome data from a limited set of conditions or tissues to improve annotations. Next-generation sequencing, new scalable experimental techniques and functional assays will enable more comprehensive sampling and more accurate prediction of function for individual genes.

One example of a powerful and scalable functional genomic approach that the DOE JGI next-generation genome center will employ is transposon-based saturation mutagenesis of culturable microbes. In this experimental paradigm, transposons are delivered into microbial genomes at nearly base-pair resolution so that a virtually saturated population of mutants can be cultured and sequenced to determine what genes are essential under any given condition (vital genes for any process never show a transposon insertion due to selective constraint). By coupling this mutagenesis methodology with automation and next-generation sequencing, functional annotation studies can be performed under many environmental conditions. It is expected that such screens will increase our ability to assign function to the large number of orphan genes presently lacking meaningful annotation. While studies of this type have been performed on a small scale, the DOE JGI with its unique capabilities in automation, sequencing and analysis is positioned to do such studies with users at a massive scale, leading to dramatic increases in the percentage of annotated genes in an organism's genome.

PILLAR 2: BIOLOGICAL DATA INTERPRETATION

BIOLOGICAL DATA INTERPRETATION – GOALS

- Maintain and expand integrated genomic data management systems as a foundation for biological data interpretation
- Scale high-throughput structural and functional annotation
- Perform deep annotation of critical protein families, non-coding genes and gene networks
- Expand genomic methods and tools for inferring function from sequence
- Strengthen strategic partnerships with supercomputer centers at DOE National Laboratories
- Establish collaborations with expert computer science and applied mathematics groups for developing scalable and efficient data processing tools for high performance computing environments
- Deploy data fusion tools for improving the quality of annotations and limiting the size of datasets provided for analysis

BIOLOGICAL DATA INTERPRETATION – STRETCH GOALS

- Develop tools to analyze, simulate and correctly predict the impact of environmental change on complex biological processes in individual organisms or entire communities of organisms
- Real-time assembly and annotation of genomes from simple and complex environments
- Accurately predict the complete metabolome of single organisms and complex communities
- Interpret the metagenomes of complex communities and their response to natural and man-made perturbations

BIOLOGICAL DATA INTERPRETATION – SCIENCE STRATEGY

- Maintain and expand DOE JGI's sequence data interpretation pipelines and integrated genomic data management
- Develop and implement robust sequence assembly algorithms for complex eukaryotic organisms and metagenomic communities
- Develop scalable metabolic reconstruction and phenotype prediction tools
- Develop scalable software optimized for performance in supercomputing environments

KEY:

- ■ Advance existing capabilities
- ■ Expand nascent capabilities
- ■ Develop new capabilities

As genomic datasets increase in scale and complexity, their systematic biological interpretation will become increasingly important to the DOE JGI's effort in support of scientific studies.

The genomic sequence data interpretation process initially involves assembling discrete sequence reads into contiguous sequence scaffolds, identifying protein and RNA genes, and characterizing genes. However, the biological functions of many genes that are currently being discovered in complete genomes or metagenomes remain unknown. The DOE JGI will address the functional characterization of novel genes by employing both high-throughput functional genomic approaches, as well as advanced computational methods to reveal hidden homologies and assign potential functions. Together these approaches will contribute to the development of the predictive models of gene function that will be required if users of the DOE JGI are to successfully apply biology-based approaches to the challenges posed by energy and environmental issues.

Biological systems are not static, but vary in time and space in ways that are poorly understood. A fundamental understanding of this variation will provide important clues for how to manipulate these systems to further DOE mission goals. The exponential growth in the number and size of genome and metagenome sequence datasets poses computational, data management and analytical challenges for their comprehensive biological interpretation. Embracing parallel computing will be necessary, but not sufficient to solve these challenges. Significant investment in computational algorithms will be needed both to interpret data and to guide subsequent experiments. The DOE JGI next-generation genome center will position itself to excel in approaches that leverage both advanced molecular methods and computational techniques, in an integrated fashion that will be available at no other facility.

Computational analysis capabilities that will be required by the DOE JGI next-generation genome center are elaborated below, along with the associated scientific goals and computational challenges.

A. ASSEMBLING FUTURE SEQUENCE DATASETS

The increase in raw sequencing capacity projected over the next decade at the DOE JGI is complicated by the expected diversity of sequence types. We are likely to see a mix of deep and accurate but short sequences as well as longer but less accurate single molecule-based sequences. We also anticipate the emergence of complex combinatorial strategies for pooling inputs that take advantage of the DOE JGI's projected scaling in handling ever-larger numbers of samples across diverse conditions. Raw sequences from diverse and ever-changing technologies must be combined into genic and chromosomal reference sequences and transcriptomes to be useful for downstream analytical, synthetic biological and other functional genomic efforts and will represent daunting algorithmic and computing challenges. Over the next few years, the computational assembly of currently available short-read datasets will increasingly become a bottleneck for complex genomes and metagenomes, since current approaches tax existing computing infrastructure in a way that is not scalable.

A major algorithmic and computational goal for the DOE JGI next-generation genome center is therefore the development and implementation of new strategies for combining diverse new sequence types into high-throughput assemblies of complex genomes and metagenomes. This computational capability goes hand-in-hand with the development of future sequencing approaches, since effective strategies for combining next-generation genetic mapping, clone pooling, chromosome sorting, and other novel technologies will depend on the detailed characteristics of each new data type. Taking full advantage of new data collection methods will require close collaboration between computational scientists and technologists at the DOE JGI to model the statistical properties of new sequencing datasets (error rates, sampling frequencies, etc.) and to develop, test, and implement new strategies for combining diverse data sets in an optimal manner for diverse user projects.

The DOE JGI's focus on genomics relevant to energy and the environment brings unique difficulties that will not be faced by other sequencing groups. For example, the strongest candidates for biofuel feedstocks (switchgrass, miscanthus, sugarcane, and other related species) are outbred grasses with significant intra-species genomic variation. In many cases, they are also polyploid, with multiple nearly identical but functionally distinct copies of each chromosome. Similarly, deep metagenome datasets harbor intra-species variation whose functional consequences are poorly understood.

The DOE JGI will develop new methods to represent polyploid genomes, microbial "pan-genomes," and metagenomes in a useful manner. These representations will identify a common core of a set of related genomes, along with a catalog of sequence differences between individuals that may include variable gene content present in some strains or samples of a species. In addition to organizing genomic information about a species, pan-genome representations will simplify data storage and retrieval, significantly reduce downstream computations, and suggest intuitive approaches for working with large datasets.

B. FUNCTION DISCOVERY AND ANNOTATION

Over the past decade, gene finding algorithms have matured to the point that bacterial and archaeal genes can be predicted from high quality sequence with great fidelity. Prediction of eukaryotic genes, including alternate-spliced forms, remains challenging, often relying on empirical transcriptome data as training sets. Annotation of predicted genes begins with comparison of sequences to genes and proteins of known function, but a significant fraction of genes predicted in organisms from the environment are either not similar to previously described genes or are similar to genes that themselves are of unknown function. This problem is compounded when sequences are incomplete or contain errors as is often the case with metagenome assemblies. If complete functional annotation is to be provided, a variety of new computational methods must be developed to improve functional predictions and place genes within metabolic pathways and gene networks.

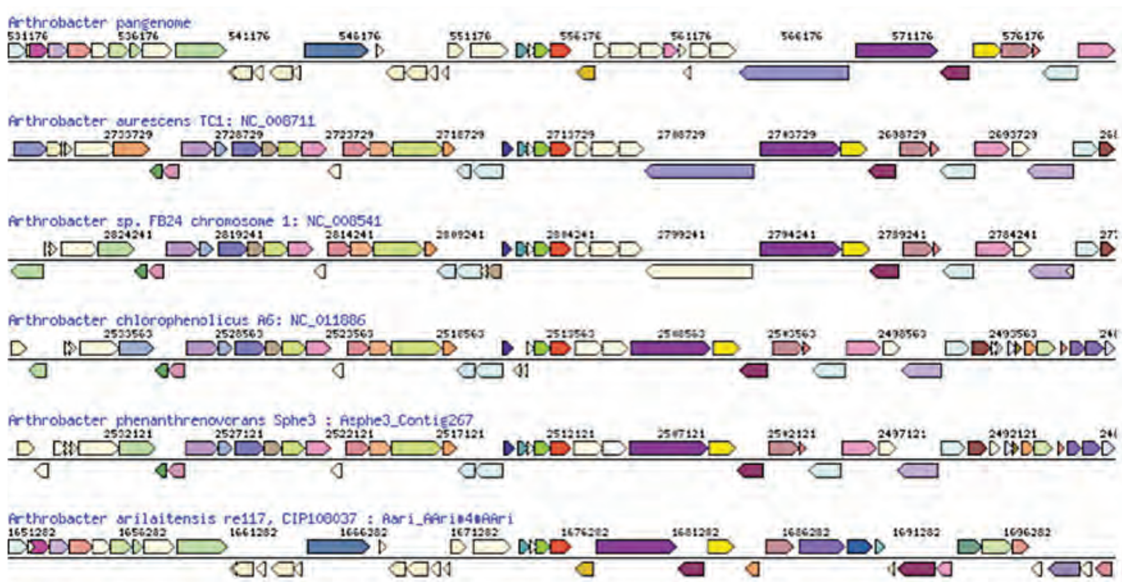
1. SCALING HIGH-THROUGHPUT STRUCTURAL AND FUNCTIONAL ANNOTATION

The structural and functional annotation of genes encoded by genomic and metagenomic assemblies and raw datasets is essential to their interpretation. Current methods of structural annotation of genes take advantage of comparative genomics to identify coding sequences in microbes and exons in eukaryotes, and integrate homology and deep transcriptome data. Sequence similarity between genes, their phylogenetic distribution, and comparison to functionally curated sets of motifs, domains, and sequence profiles are the current basis for inferring the function of each gene. These still-evolving methods, however, will have prohibitive computational costs if implemented on the scale needed over the next 5-10 years. DOE JGI computational scientists, working in conjunction with external partners, will develop algorithms that scale in response to exponential growth in the volume of raw sequence. High-performance computing, including cloud architectures under development at DOE supercomputing centers and elsewhere, will be harnessed with appropriately optimized algorithms.

In the coming decade, the rapid growth in sequence capacity will allow the development and expansion of new approaches to functional annotation based on sequence comparisons. Conservation at the level of nucleotide and peptide similarity, genomic context (e.g., operons, synteny), phylogenetic distribution, and domain linkage are currently used to transfer functional annotations on a modest scale. In the near future we will clearly need newer scalable algorithms that take advantage of the availability of hundreds of billions of genes from diverse genome and metagenome sequencing efforts. Individual genomes will give way to hundreds or thousands of closely related cultivars or strains, and the composite pan- or multi-genome will be the basis for annotation and analysis. In the circumstance where each genome has a cloud of sequenced relatives, along with related data from metagenomes, the older paradigms will need to be abandoned in favor of more efficient and biologically informed computational approaches.

2. FUNCTIONAL ANNOTATION OF NOVEL GENES AND GENOMES

A significant fraction of genes in sequenced genomes have no known function. Some may have sequence similarity to genes in other species, and represent entire gene families without assigned function. Others have little sequence similarity to other genes, having diverged from their homologs.



Pangenome representation of a set of organisms belonging to the genus *Arthrobacter.* The first line corresponds to the computed pangenome. A pangenome consists of the core part of a species (i.e. the genes present in all of the sequenced strains or of all samples of a microbial community) and the variable part (the genes present in some but not all of the strains or samples). Colored arrows indicate equivalog genes. The pangenome offers a more compact data size view while maintaining the gene context information.

A persistent challenge in the functional annotation of genes and genomes has been the ongoing discovery of such novel genes. The absence of detectable sequence similarity to genes of known function stymies the transfer of functional information across species. Even with scale-up of current methods, the functions of these genes will remain mysterious without development of new methods and approaches.

The DOE JGI will take a multi-pronged approach to characterizing the functional roles of such genes, combining "data production" activities including directed synthesis, functional assays, genetic analyses of variants, and custom computational analysis. New algorithms for "deep" sequence comparisons will be developed that transcend the currently available BLAST and Hidden Markov Model-based approaches. With thousands of homologous sequences, methods that integrate structural alignment methods will go beyond linear sequence comparisons towards approaches that explore the three-dimensional structure of peptides *in silico* to complement high-throughput assays produced by the DOE JGI and structural genomics efforts at national user facilities managed by the DOE and other agencies. We expect that these methods will take advantage of deep collections of metagenomes to identify divergent homologs and in this way develop statistical models for new classes of genes. In deeply sequencing extreme (or even conventional) environments, we anticipate that previously unknown clades will be discovered. Understanding such novel branches of life, initially identified through sequence, will likely rely heavily on new algorithms for functional inference, as well as expert application of phylogenetic methods.

3. DEEP ANNOTATION OF SELECTED GENE FAMILIES

For applications envisioned by the DOE JGI, including the design and synthesis of genes and genomes that perform useful functions relevant to energy and the environment, it will be essential to develop the capability to predict and redesign/engineer enzymes to alter substrate preferences, optimal pH, temperature, etc. These properties cannot currently be predicted from sequence alone. The function of a gene currently can only crudely be assigned based on sequence similarity, domain composition, and functional sequence motifs. For example, we may identify a gene as a cellulose hydrolase by sequence comparison, recognizing a generic domain, but the specific enzymatic properties of this gene remain unknown. DOE JGI users will require far more specific predictions of functional activity.

Deep collection of sequences from judiciously chosen genomes, metagenomes, and transcriptomes, provide a broad sampling of the "sequence space" of a gene family of interest. Computational analysis of this sequence set will enable the design of candidate genes/peptides that can be synthesized in bulk and subjected to high-throughput functional assays. Specific sequences at a known active site can be systematically varied and tested. DOE JGI users will want to design systems for use under a wide range of conditions. For example, the deconstruction of biomass, the reengineering of microbial and fungal systems for enzyme production, and the optimization of cell wall chemistries of biofuel feedstocks, will each require modifications to be made to collections of enzymes at structural positions far from the active site. Thus, the DOE JGI next-generation genome center will require computational structural biology, working in an interactive and iterative mode with high-throughput functional testing for specific bioenergy and bioremediation applications.

4. INTERPRETING METAGENOME DIVERSITY AND DYNAMICS

By scaling current methods we can now enumerate the likely functional content of a microbial community, in the form of a "catalog of genes." Some fraction of these genes will be of unknown function, including some without currently recognizable sequence similarity to known genes. Such a long list of relatively easily detected component genes is the natural first target, and can in principle be accomplished by an extension of the conventional annotation of microbial genomes. The current scale of metagenomics, however, already introduces significant algorithmic and computational challenges.

Experiments in the near future will involve analysis of the response of microbial communities to natural environmental variation and experimental perturbation; the dependence of microbial communities on the genotype and phenotypic state of their host plant; and the sampling of metagenomes across hundreds of diverse environmental conditions. This variety is layered on top of the ever-increasing throughput of sequencing instruments, allowing deeper sampling of each metagenome. Directly comparing each metagenome sequence read to all others is not computationally feasible and will likely remain so. Even comparing metagenome sequence fragments to gene families extracted from current microbial genome collections, which fails to capture the genetic novelty in metagenome data, already taxes the limits of currently available high-performance computing infrastructure.

Methods for reducing the size of the data to be compared, including both the metagenome query data (e.g. assembly) and the target databases (e.g., collapsing related genomes into "pangenomes") are critical as are more efficient search tools and techniques.

At the same time, a simple list of potential genes and functions defies meaningful interpretation, requiring effective methods for comparison and visualization to make sense of the high-dimensional data and infer higher-order laws governing community behavior. Clearly experimental data, including high-throughput single cell genomics and direct functional analysis of novel gene families will contribute greatly to the desired functional understanding, but new analytical tools are still needed to effectively integrate these data. Ideally, such tools might allow preliminary data to guide in-depth analysis, for example by identifying a set of "focal" genes or pathways important in a given environment; this could help circumvent the obstacles posed by annotating every sequenced base. To move beyond the "catalog" state, metagenomic science will thus require two-way interactions between computational analysis and new experimental studies designed to test proposed higher-order laws governing microbial communities. Collaborative relationships between the DOE JGI, other informatics resources such as the DOE Knowledgebase and DOE's National Laboratory partners will be developed to respond to this important challenge for DOE JGI users over the next decade. Creative approaches for efficiently utilizing available resources will need to be taken to avoid the computational calamity predicted from the relative increase in the rate of advances in DNA sequencing compared to that of computing.

5. FUNCTIONAL ANNOTATION OF GENES AND NETWORKS

The DOE JGI next-generation genome center, in concert with the DOE Knowledgebase, will measure, predict, and ultimately redesign organisms and communities of organisms for deployment in service of the DOE mission. These synthetic biology efforts will be informed by integrated analyses of organismal and community responses to biotic and abiotic challenges. Microbial systems will be the initial proving ground for these approaches, as computational methods are developed to integrate transcriptomic, proteomic, and metabolomic datasets under controlled conditions. For multicellular eukaryotes, as well as for microbial communities, the complexity is increased by the presence of diverse cell types or species, each interacting with one another through poorly understood direct and indirect signals. For example, over the next 5 years, the DOE JGI Plant Gene Atlas project will measure the expression levels of tens of thousands of genes in flagship plant genomes through RNAseq and other high-dynamic-range expression profiling datasets (replicated controlled experiments across tissues, single-cell types within tissues, varying conditions, and genotypes). DOE JGI computational scientists, in conjunction with collaborators representing the user communities for the flagship genomes, will be the primary analysts of these datasets, feeding high quality expression analyses into the DOE Knowledgebase. Over the next decade, we envision producing and analyzing other large-scale profiling methods that will provide a rich quantitative characterization of biological systems critical to the DOE mission.

In concert with the modeling community and the DOE Knowledgebase, DOE JGI scientists and its partners will develop algorithms for the analysis of such high-dimensional datasets, with the goal of extracting predictive network datasets for key plants, fungi, microbes, and microbial communities relevant for DOE. The DOE JGI will develop core datasets that enable the correlative analyses that will ultimately lead to predictive models of system response to perturbations. Bidirectional partnership with the DOE Knowledgebase will be essential to developing such models.

The tight coupling of these predictive models with the synthetic biology capacity of the DOE JGI will allow users to engineer new systems that follow the general principles learned from natural systems, but which execute novel functions tailored to particular applications. It is unlikely that there will be a linear path from comprehensive reference datasets to predictive models to newly designed systems that work "out of the box." Rather, DOE JGI users, and DOE JGI computational scientists and technologists will need to work closely to iterate between modeling and experimentation/testing to expose the principles that underlie robust networks, and provide more robust predictions that can be lifted from models to systems "in the field."

6. NATURAL VARIATION AND GENOME-ASSISTED BREEDING

Genetic variation underlies heritable phenotypic variation, and provides the raw material for the improvement of key bioenergy and bioremediation phenotypes through directed breeding. Within the next few years DOE JGI and its collaborators will develop comprehensive catalogs of genetic variation for "flagship" systems of DOE relevance. These will require computational analysis that will need to become standardized.

The larger computational and analytical challenge, however, is to relate genotype to phenotype. By crossing extreme phenotypes, quantitative trait loci can be found, and candidate gene variants identified; by characterizing functionally divergent populations, genome-wide-association methods can be applied, with suitable corrections for population structure. Thus the DOE JGI must provide users with expertise in statistical and population genetics. Statistical models that predict phenotype from genotype can then be used for genome-assisted breeding, which will move from marker-assisted methods involving a few loci to "genomic selection" methods that iteratively adjust genotype-phenotype models in real time, integrating new information from each accelerated breeding cycle.

An almost completely untouched problem that DOE JGI researchers and users will confront is the modeling of metagenome-to-phenotype relationships. Proper analysis of such datasets will require statistical and population genetic expertise, coupled to high performance computing to handle the unprecedented scale of the analyses necessary to make use of these complex datasets.

C. COMPUTING REQUIREMENTS

Raw sequence data from genomes and metagenomes are transformed into biologically meaningful information using computational tools and pipelines. A comprehensive sequence data interpretation process employs the integrated data context of an expanding universe of genome and metagenome sequence datasets, and involves incorporation of complementary 'omics' technologies for validating the coherence of biological information. Data interpretation is also inherently iterative, since repeating one or several of the processing stages in the presence of ever-growing datasets gradually improves the breadth and depth of biological information.

Sequence data interpretation and integration processes must be scalable to cope with the increase in the rate of sequencing of genomes and metagenomes, the size of metagenome datasets generated using new sequencing platforms, and the diversity of 'omics' datasets. The estimated size of datasets generated with new genome sequence technology platforms are expected to grow faster than the computing resources available to the DOE JGI. Addressing this challenge requires leveraging additional computing resources and importantly developing scalable and efficient data processing tools.

1. LEVERAGING COMPUTING RESOURCES

Analysis of high throughput sequence data in an 'omics' context requires High Performance Computing (HPC) capabilities set in a High Throughput Computing (HTC) environment. The DOE JGI presently relies on Lawrence Berkeley National Lab's National Energy Research Scientific Computing Center (NERSC) for supporting its High Performance & Throughput Computing (HPTC) needs, with a compute cluster and large capacity distributed file system maintained by NERSC at its core. Additional computing resources will be sought through partnerships with other DOE Leadership Computing Facilities, with surge computing needs provided via on-demand access to cloud computing resources. Leveraging HPC platforms at NERSC and other DOE Leadership Computing Facilities will require refactoring sequence data processing and integration pipelines to run efficiently on these platforms.

2. DATA PROCESSING EFFICIENCY AND SCALABILITY

Next-generation sequencers are expected to generate petabases of sequence data in a matter of days in the next decade. Keeping up with this increase in the size of sequence datasets will be crucial for scientific advances. Accordingly, the efficiency and scalability of data processing tools and pipelines need to be continuously improved.

3. REENGINEER ROUTINE DATA PROCESSING TASKS

Data processing tasks, such as sequencing quality control for base calling, detection of contamination, sequence alignment, assembly, and gene prediction, will need to be reengineered to run efficiently as part of automated pipelines on existing and emerging exascale supercomputers, with minimum or no supervision from data analysts. Recently, several large-scale assembly tasks capable of handling terabases of sequence data have been implemented using MPI (Message Passing Interface standard) to run on DOE supercomputers. In the future, additional data processing tasks will need to be parallelized to form an efficient sequence toolkit that can be then combined in pipelines targeting specific scientific goals.

4. CUSTOMIZATION OF DATA PROCESSING TASKS

Customization of data processing tasks is driven by unique aspects of specific studies and may range from minor changes, such as adjusting parameters or execution order or tools, to major changes involving development of new analytical capabilities or approaches. Minor customization of data processing tasks involving adjustment of existing tools and pipelines need to avoid time consuming software development, and will be explored in the framework of the Hadoop programming model for data parallel applications based on the Map/Reduce paradigm. New analysis challenges will likely require special purpose computing capabilities, such as high-performance computing systems with petabytes of RAM that could be accessed through partnerships with DOE supercomputer centers and/or commercial providers.

5. ALGORITHMIC IMPROVEMENTS

While computing resources at DOE Leadership Computing Facilities can be leveraged to address DOE JGI's computing needs in the next 1-2 years, these will not be sufficient to keep pace with the growth in sequence data generation expected over the next decade. Consequently, new data processing approaches and methods need to be devised with the goal of reducing the amount of computation required for the biological interpretation of genome and metagenome sequence datasets. For example, data processing may shift from exhaustive to targeted interpretation, such as focusing on key classes of enzymes for biomass degradation and conversion, or regulatory networks for cell wall metabolism.

For its computing requirements, the DOE JGI next-generation genome center will strengthen its strategic partnerships with supercomputing centers as providers of DOE JGI's computing needs, enhance DOE JGI internal capabilities for parallel algorithm development, and pursue collaborations with expert computer science and applied mathematics groups for the development of data processing methods that are scalable and perform efficiently on high performance computing infrastructures.

6. TRAINING IN PROGRAMMING FOR THE HIGH PERFORMANCE COMPUTING ENVIRONMENT

A crucial bottleneck in integration of genomics and high performance computing is the limited number of computer scientists and engineers that are conversant in the two disparate fields. Extensive exchanges and cross-training of individuals working in these areas will be developed through tutorials, internships, and formal exchanges between DOE High Performance Computing facilities and the DOE JGI. The ultimate goal is to develop a cadre of individuals skilled and knowledgeable in both genomics and high performance computing who can in turn use this expertise to enable users of the DOE JGI to solve cutting-edge problems requiring the analysis of large genomic datasets.

PILLAR 3: USER INTERACTIONS

The DOE JGI is first and foremost a user facility. The users of the DOE JGI are varied and represented by:

- scientists who gain access to sequencing, analysis and experimental capabilities available at the DOE JGI and its partners through a peer review process
- communities of scientists and bioinformaticians who draw upon the display and analysis of data generated by the DOE JGI, as well as other relevant data that informs DOE JGI data
- scientists educated by the DOE JGI in the use of genomics to solve energy and environmental problems
- community resources, such as the DOE Knowledgebase, that access JGI data, process them, link them to other data sets and provide a general integrated framework for bioinformaticists and biologists to model and exploit these data

USER INTERACTIONS – GOALS

- Interactive Data Platforms: Design and implement interactive data platforms that enable users to address questions across multiple levels of complexity (genome, cell, tissue, organism, community, environment, life-cycle) using DOE JGI sequence and functional data
- Customized User-Data Interactions: Develop common software frameworks for user portals, data sites, and application programming interfaces to support consistent and customized user-data interactions
- Integrated 'Omics': Integrate world-wide genomic and 'omics' data with DOE JGI data
- Standards: Develop standards and interfaces supporting DOE JGI data exchange and interoperability with external user tools and the scientific community
- Training the Next Generation of Genomic Users: Develop educational and training programs to enable future DOE JGI users to exploit DOE JGI information
- On-site User Science: Develop programs for hosting users at the DOE JGI to carry out complex genomic experimentation and analysis supported by DOE JGI staff and provide users with ready access to needed capabilities provided by DOE JGI partners
- Organize Mission Oriented Communities: Develop new active role in bringing communities together, coordinating and providing them with data and capabilities to address DOE's most pressing challenges

USER INTERACTIONS – STRETCH GOALS

- Facilitate user-driven multidimensional data analysis enabling users to navigate across sequence and functional data types (genomic, transcriptomic, proteomic and metabolomic) and domains of life
- Enable users to create personal portals with customized components tailored to their own research goals
- Extend data mining and analysis tools into the realms of system organization and behavior
- Provide efficient methods for the real-time interactive analysis of user data in the context of DOE JGI's integrated data systems
- Enable a majority of scientists working in energy and the environmental sciences to access and exploit genomic data and advanced 'omic' analysis capabilities at the DOE JGI and its partners

USER INTERACTIONS – SCIENCE STRATEGY

- Develop user interfaces and computational infrastructure to enable fast, accurate and informative data access and retrieval, data mining and analysis
- Expand DOE JGI educational workshops, tutorials and undergraduate education program
- Develop community-accepted standards for metadata associated with single cells, isolate genomes and metagenomes
- Formalize DOE JGI's role in organizing user communities to facilitate access to DOE JGI capabilities and sharing of strategies, analyses and external data
- Develop active visiting scientist and outreach programs to bring investigators to the DOE JGI to carry out high-impact studies
- Develop seamless interfaces with the DOE Knowledgebase and other community resources
- Develop a common software framework for user portals that supports both consistency and customization of user-data interactions

KEY:

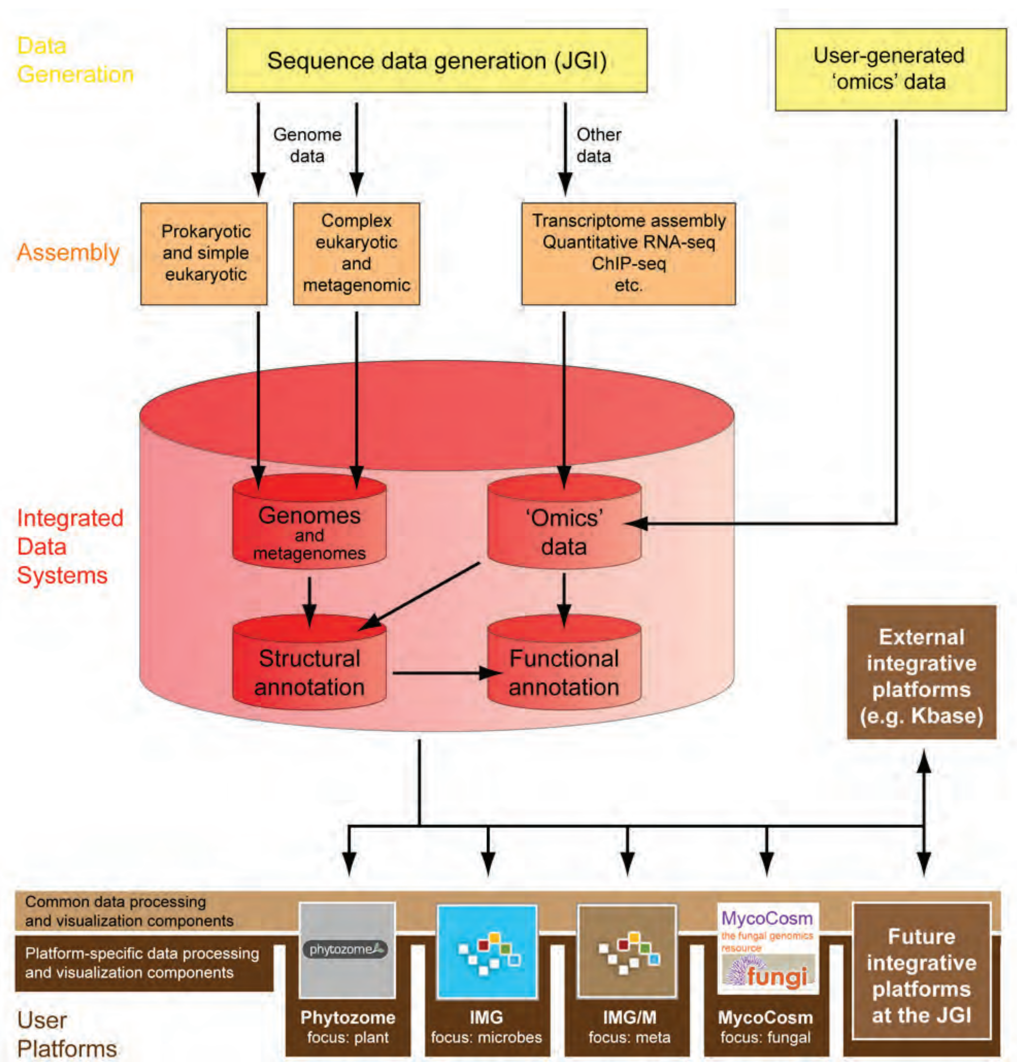
- Advance existing capabilities
- Expand nascent capabilities
- Develop new capabilities

A. DATA PLATFORM-BASED USER INTERACTIONS

OVERVIEW

Presently the JGI user communities are organized around the DOE JGI Science Programs (Microbial, Metagenomic, Fungal and Plant Genomics) and the programs' interactive data platforms (IMG, IMG/M, Mycocosm and Phytozome). These systems provide both internal and external users with access to DOE JGI-generated and -processed data and results of analysis pipelines, in a framework that supports data searching, visualization and additional user-driven analysis.

In the future, as the DOE JGI evolves into a next-generation genome center, an increasing number of the scientific problems it addresses will cross the current program boundaries. In response to this development, the DOE JGI's integrated data platforms will be extended and transformed to improve their utility as a tool for users to explore mixed environmental systems. Horizontal integration between existing program portals will be achieved by the development of shared data models, a common organization of data and shared tools for analysis and visualization.



To address the increasing need for integrative studies that span all kingdoms of life and complex environmental samples, data analysis systems and user-accessible data platforms will be increasingly integrated at the DOE JGI next-generation genome center. This integration will also facilitate data exchange and interoperability with external advanced integrative data platforms.

1. DATA MINING AND ANALYSIS: QUERYING, MINING AND VISUALIZING MULTIDIMENSIONAL DATA

Unprecedented data growth poses new challenges for data analysis and presentation. In the next few years, hundreds of thousands of genomes, millions of samples and billions of genes will be sequenced and become available for analysis by the community. Integration of related data from different ontological domains (e.g., functional, metabolomic, phylogenetic, environmental, phenotypic, geographical) will create even larger volumes and higher levels of complexity. Presenting these multidimensional data to users will require novel methods for querying, mining and visualization. The types of queries that are currently widely used will generate results that are larger by orders of magnitude. To present such enormous datasets in manageable form, new and original tools will be built that operate through categorizing, organizing and filtering selected data sets on the fly, alleviating the need to visualize vast lists of objects. Such instant categorization, organizing and filtering will comprise an entirely new approach to data mining. These new visualization interfaces will also enable seamless integration of multiple ontological perspectives, and allow for interactive switching between them.

2. COMMON SOFTWARE FRAMEWORK FOR USER PORTAL CONSISTENCY AND CUSTOMIZATION

The DOE JGI's integrated data platforms are presently optimized to serve the analysis and visualization needs of their respective scientific programs and associated users. In the future the DOE JGI will develop a common software framework that will support multiple programs and enable cross-program tool development. Moving to a common platform software framework will involve:

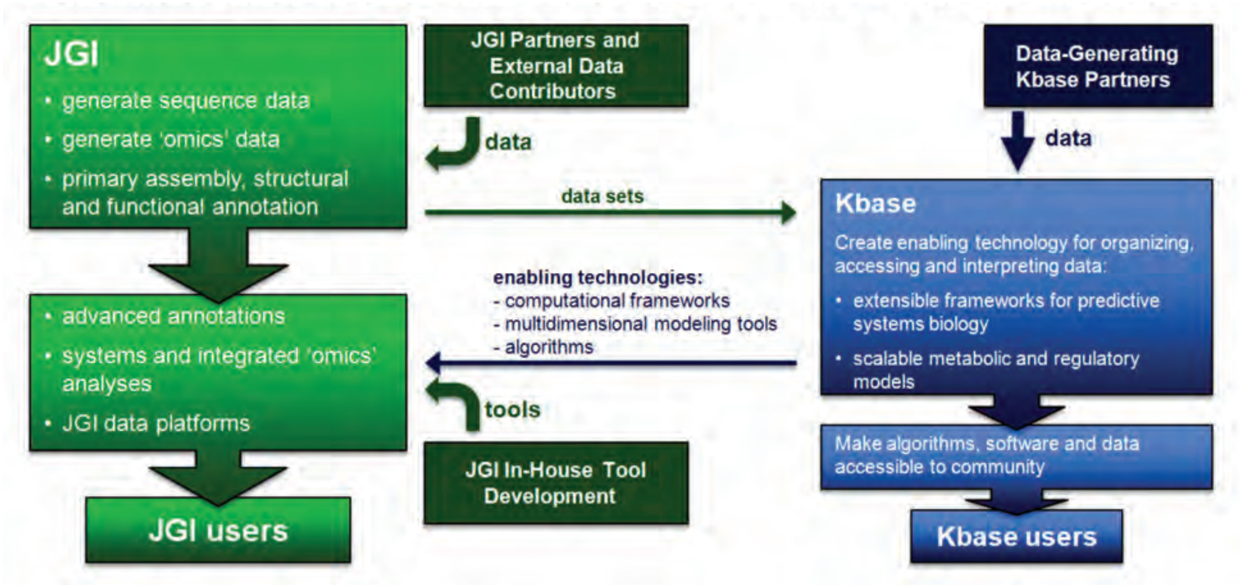
- Providing a common set of visualization and analysis components that can be deployed in multiple platforms, minimizing development effort and user training, and freeing resources to work on new capabilities
- Developing an intermediate data model to insulate platform component design from challenges related to backend data storage models
- myIDP: provide default integrated data platform configurations (portal layout) guided by current program standards, that allow end users to customize the components, layout and data types included

3. ENABLING CUSTOMIZED RETRIEVAL OF SUBSETS OF ALL DOE JGI DATA

Present DOE JGI data dissemination strategies are dominated by pre-computed, non-customizable processed data and analysis results sets that are distributed to users via interactive web-enabled data browsers. In addition, custom data sets are provided to users by program teams wherever possible. These web-enabled data sets are almost always kingdom-specific, and their construction and iterative refinement is a labor-intensive manual process. In the future, the DOE JGI will increasingly provide tools that enable users to directly construct customized, bulk-scale data sets, in a variety of standardized genomic formats, from the full range of processed, analyzed data and external data available through DOE JGI systems. Limited versions of this type of capability are currently implemented in the open source BioMart and Intermine projects, which will serve as useful testbeds for the development of such systems at the DOE JGI. A major focus of the development in this area will be the improvement of loading and retrieval performance issues that are expected to occur as data sets increase massively in size.

4. ENABLE PROGRAMMATIC ACCESS TO DOE JGI DATA

In the past, the DOE JGI focused on providing effective data access and analysis tools to end users, whereas external computational systems and tool developers were less effectively served, due to the lack of uniform, comprehensive Application Programming Interfaces (APIs). Currently, APIs are only available for some of the integrated data platforms at the DOE JGI, with Mycocosm and the BioMart component of Phytozome supporting access to subsets of their processed data. Providing well-documented, complete APIs with reference implementations will allow the data platforms at the DOE JGI to seamlessly connect to external genomic cyberinfrastructure (such as the DOE Knowledgebase and iPlant) to take advantage of tools, pipelines and computational resources located outside of the DOE JGI. This will also enable moderately sophisticated end users to develop custom analyses of DOE JGI data without the need to understand or interact with the data storage model at the DOE JGI. Finally, it will allow outside developers to produce analysis and visualization components and tools that can be contributed back to the general community of DOE JGI data users, lowering the access and analysis barrier for all users.



Synergistic Interactions between the DOE JGI next-generation genome center and the DOE Knowledgebase.

B. DIRECT DOE JGI USER INTERACTIONS

The DOE JGI is the only DOE genomics user facility and its interactions with users occur in a variety of ways, some of which are similar to user interactions at other DOE user facilities, while others are unique to the DOE JGI. As the DOE JGI evolves into a next-generation genome center, with a variety of capabilities to offer users in addition to high-throughput sequencing, it moves into a new phase as a user facility. The multiple specialized capabilities being applied to the complex projects of the future will result in increasing numbers of users needing to spend time at the DOE JGI and interacting with the skilled staff of experimentalists and bioinformaticists to fully exploit its capabilities. Visiting scientist programs will be developed to facilitate these new interactions.

1. PROVIDING USER ACCESS TO CAPABILITIES OF THE DOE JGI PARTNERS

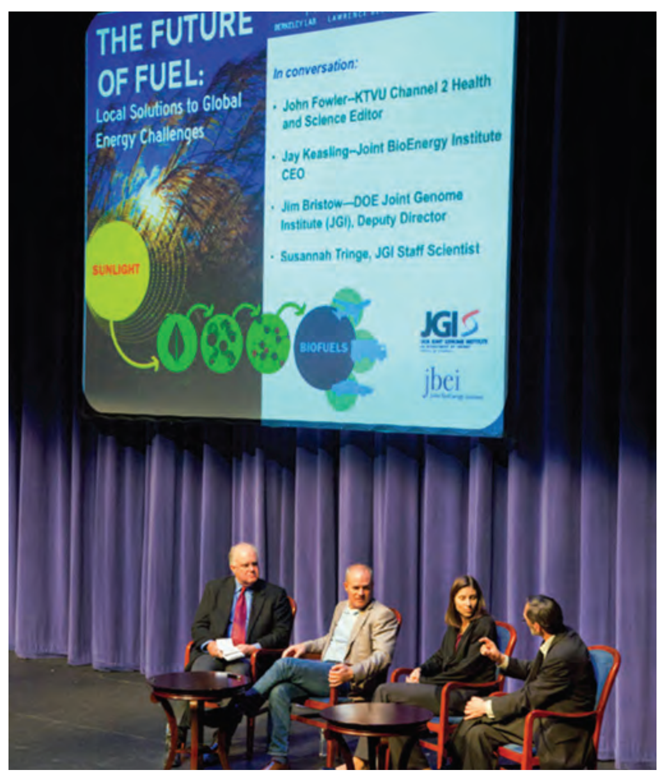
Since its beginning in 1999, the DOE JGI was historically a collaborative partnership among several DOE National Laboratories. Funding went directly to different partner institutions with the DOE JGI Director coordinating the scientific activities but not the funding of these institutions. This partnership has evolved and the DOE JGI Director now has the flexibility to directly move resources to external partners based on the institute's scientific needs. A core philosophy of the DOE JGI as it goes forward is that its capabilities will need to continuously evolve to ensure that it remains a state-of-the-art next-generation genome center. New capabilities that will be sought from partners are those where highly specialized and/or mature versions of needed capabilities already exist, obviating the need for them to be developed by the DOE JGI. This will range from data generation capabilities (proteomics from an institution such as PNNL's Environmental Molecular Sciences Laboratory) to analysis-related capabilities (access to appropriate software and high performance computing infrastructures as well as the DOE Knowledgebase). The DOE JGI will develop interfaces to enable users to access a diversity of DOE JGI partner capabilities to advance their science.

Access to extended computational resources and experimental capabilities will be made available to JGI users through an evolving set of partnerships.

2. ORGANIZING MISSION-ORIENTED USER COMMUNITIES

Since its inception as a user facility in 2004, DOE JGI has been helping user communities organize into cohesive groups to accomplish specific project goals. Organization is essential for communities in planning their projects, delivering materials in a timely fashion, carrying out specific analyses, sharing of data and publishing their work. As genomic projects become larger and more complex, organization of user communities becomes an increasingly critical prerequisite for successful completion of projects. As such, DOE JGI will formalize this organizational role as part of the next-generation sequencing center concept.

This effort will begin at the planning stages of large-scale genomic projects. Community leaders will be encouraged to engage DOE JGI scientists and technical staff in planning the scale, scope and strategy of their projects. In addition, early organizational efforts will be directed at providing access to newly developed advanced DOE JGI technologies of which the user community may not be aware. Once projects have been approved by external peer review, DOE JGI will regularly convene efforts to encourage early data sharing. DOE JGI will continue to work with the community throughout the project life cycle to help recruit investigators and analysts with expertise required for specific projects.



Left to Right: KTVU Channel 2 Health and Science Editor John Fowler, JBEI CEO Jay Keasling, DOE JGI Staff Scientist Susannah Tringe, and DOE JGI Deputy Director Jim Bristow, at the Lesher Center in Walnut Creek. (Credit: Roy Kaltschmidt, LBNL).

C. TRAINING THE NEXT GENERATION OF DOE JGI USERS

The evolution of the DOE JGI from a sequence production facility to a next-generation genome center with advanced capabilities will be accompanied by a focused effort in training future users of the DOE JGI. Education and workforce training has been identified as a long-term strategic goal and a Grand Challenge for biological and environmental research (2010 DOE *Grand Challenges for Biological and Environmental Research: A Long-Term Vision*). In response to the specific recommendations and as part of its evolution into a next-generation genome center, the DOE JGI will engage in efforts in several areas.

1. ENGAGING SCIENCE EDUCATORS

Education Programs at the DOE JGI will include training undergraduate and graduate faculty in how to incorporate genomics and bioinformatics into the life science curriculum. Faculty development is a key part of helping the U.S. to maintain its competitive edge in science and technology; keeping faculty current with the latest in genomics-based research is an ongoing challenge. The DOE JGI's faculty development efforts will be accomplished through workshops at the DOE JGI and national scientific organization meeting such as the American Society of Microbiology. These efforts will lead to the formation of a large network of collaborative faculty educators nationwide that will be an important part of the DOE JGI User Community. They will play a key role in the evolution of the DOE JGI Education Program through expansions into functional genomics. Devising a systematic approach to functional genomics at the undergraduate level will allow students to take their bioinformatics-generated hypotheses and test them in the wet lab.

2. COLLABORATIVE TEACHING PROGRAMS

The DOE JGI's Education Program will work with faculty associated with science training to build tools and systematic resources to enable undergraduate functional genomics. The Education Program at the DOE JGI is committed to keeping pace with advances in genomics-based research. The progression from bioinformatics, to functional genomics and ultimately synthetic biology will enable the DOE JGI to continue to contribute to the training of the next generation of scientists in the latest methods of DOE mission-oriented research in the life sciences. In this context, the DOE JGI will seek interactions with education experts who can help evaluate and refine the elements of the education program and study its impact on students and educators.

Engage science educators by developing collaborative teaching programs.

APPENDICES

CONTRIBUTORS AND STRATEGIC PLANNING PROCESS

Strategic planning at the DOE JGI is a continuous process, with a formal planning document generated on a tri-annual basis. This emphasis on frequent strategic planning is mandated by the extremely dynamic scientific and technological developments in the field of genomics. The process that led to the present 10-Year Strategic Vision included:

- a three-day off-site Strategic Planning Workshop with an external panel of scientific leaders in the field
- workshops and meetings focused on strategic planning with the
  - Plant User Advisory Committee
  - Fungal User Advisory Committee
  - Microbial and Metagenomic User Advisory Committee
  - DOE JGI Scientific Advisory Committee
  - DOE JGI Informatics Advisory Committee
- a town-hall session at the 2011 DOE JGI User Meeting attended by ~400 DOE JGI users, where community input was solicited on how the JGI should evolve as a user facility to best meet their scientific needs
- a three-day workshop organized by the JGI, entitled "High Performance Computing and the Needs of Genomics" that brought ~70 members of the high-performance computing community to the JGI to contribute to the JGI's strategic vision in this area

These various User Committee members, Advisory Committee members, general users and external expert scientists (totaling >500 individuals attending 8 workshops/meetings focused on DOE JGI strategic planning) provided input on the DOE-relevant science that they believed needed to occur in the next decade and the capabilities that would be required to accomplish these scientific goals. This input resulted in this document which has been reviewed and commented on by the members of the various DOE JGI User Advisory Committees, the DOE JGI Scientific Advisory Committee and the participants at the off-site Strategic Planning Retreat. Importantly, this is a living document as additional workshops will be held to derive additional input and further develop a vision for the future of the DOE JGI as a next-generation genome center.

CONTRIBUTING AUTHORS

Individuals who captured the input from the strategic planning workshops and meetings and contributed to the writing of the present draft 10-year strategic vision plan:

Harvey Bolton, Pacific Northwest National Laboratory

Siobhan Brady, University of California, Davis

James Bristow, DOE Joint Genome Institute

Bob Cottingham, Oak Ridge National Laboratory

Jeff Dangl, University of North Carolina, Chapel Hill

Jonathan Eisen, University of California, Davis

David Goodstein, DOE Joint Genome Institute

Cheryl Kerfeld, DOE Joint Genome Institute

Victor Markowitz, DOE Joint Genome Institute

Bernhard Palsson, University of California, San Diego

Len Pennacchio, DOE Joint Genome Institute

Daniel Rokhsar, DOE Joint Genome Institute

Eddy Rubin, DOE Joint Genome Institute

Jeremy Schmutz, HudsonAlpha Institute for Biotechnology/DOE Joint Genome Institute

Gary Stacey, University of Missouri

Rick Stevens, Argonne National Laboratory

Susannah Tringe, DOE Joint Genome Institute

Ray Turner, DOE Joint Genome Institute

Gerald Tuskan, Ridge National Laboratory

Axel Visel, DOE Joint Genome Institute

Tanja Woyke, DOE Joint Genome Institute

Stan Wullschleger, Oak Ridge National Laboratory

PARTICIPANTS ASILOMAR DOE JGI STRATEGIC PLANNING WORKSHOP, 2011

Harvey Bolton, Pacific Northwest National Laboratory

Siobhan Brady, University of California, Davis

Jim Bristow, DOE Joint Genome Institute

Shane Canon, National Energy Research Scientific Computing Center (NERSC)

Patrick Chain, Los Alamos National Laboratory
Bob Cottingham, Oak Ridge National Laboratory
Jeff Dangl, University of North Carolina, Chapel Hill
Jonathan Eisen, University of California, Davis
David Gilbert, DOE Joint Genome Institute
John Glass, J. Craig Venter Institute
Dan Goodstein, DOE Joint Genome Institute
Sarah Grant, University of North Carolina, Chapel Hill
Igor Grigoriev, DOE Joint Genome Institute
Cheryl Kerfeld, DOE Joint Genome Institute
Nikos Kyrpides, DOE Joint Genome Institute
Rob Knight, University of Colorado, Boulder
Victor Markowitz, DOE Joint Genome Institute
Mary Miller, DOE Joint Genome Institute
Ken Nealson, University of South Carolina
Trent Northen, Lawrence Berkeley National Laboratory
Bernhard Palsson, University of California, San Diego
Len Pennacchio, DOE Joint Genome Institute
Margaret Riley, University of Massachusetts
Dan Rokhsar, DOE Joint Genome Institute
Eddy Rubin, DOE Joint Genome Institute
Jeremy Schmutz, HudsonAlpha Institute for Biotechnology
Gary Stacey, University of Missouri
Rick Stevens, Argonne National Laboratory
Susannah Tringe, DOE Joint Genome Institute
Ray Turner, DOE Joint Genome Institute
Jerry Tuskan, Oak Ridge National Laboratory
Axel Visel, DOE Joint Genome Institute
Chia-Lin Wei, DOE Joint Genome Institute
Tanja Woyke, DOE Joint Genome Institute
Stan Wullschleger, Oak Ridge National Laboratory

Front Row (from left): Len Pennacchio, Eddy Rubin, Rick Stevens, Sarah Grant, Jeff Dangl, Jonathan Eisen, Trent Northen, Axel Visel, Ray Turner. Second Row (from left): Jim Bristow, Chia-Lin Wei, Jeremy Schmutz, Margaret Riley, Ken Nealson, Bob Cottingham, Stan Wullschleger, Tanja Woyke, Harvey Bolton, Cheryl Kerfeld, Patrick Chain. Third Row (from left): Jonathan Glass, Igor Grigoriev, Victor Markowitz, Shane Canon, Siobhan Brady, Gary Stacey, Dan Rokhsar, David Goodstein, Nikos Kyrpides, Jerry Tuskan, Susannah Tringe.

PLANT PROGRAM USER ADVISORY COMMITTEE

Jeff Dangl, University of North Carolina
Joe Ecker, The Salk Institute for Biological Studies
Eva Huala, Carnegie Institute / TAIR
Sabeeha Merchant, University of California, Los Angeles
Thomas Mitchell-Olds, Duke University
Stephen Moose, University of Illinois
Cathy Ronning, DOE BER *(ex officio)*
Gary Stacey, University of Missouri

FUNGAL PROGRAM USER ADVISORY COMMITTEE

Joan Bennett, Rutgers University
Randy Berka, Novozymes
Daniel Cullen, Forest Products Laboratory
Daniel Eastwood, Warwick University (UK)
Stephen Goodwin, Purdue University
David Hibbett, Clark University
Thomas Jeffries, USDA Forest Service Forest Products Laboratory
Gert Kema, Plant Research International (Netherlands)
Christian Kubicek, Vienna University of Technology (Austria)
Cheryl Kuske, Los Alamos National Laboratory
Francis Martin, INRA (France)
Conrad Schoch, NCBI
Joseph Spatafora, Oregon State University
Jason Stajich, UC Riverside
John Taylor, UC Berkeley
Adrian Tsang, Concordia University (Canada)
Gillian Turgeon, Cornell University
Ryta Vilgalys, Duke University

MICROBIAL GENOMICS AND METAGENOMICS PROGRAM USER ADVISORY COMMITTEE

Swapnil Chhabra, Lawrence Berkeley National Laboratory/Joint BioEnergy Institute
Cameron Currie, University of Wisconsin
Jed Fuhrman, University of Southern California
Steve Hallam, University of British Columbia
Phil Hugenholtz, University of Queensland
Javier Izquierdo, BioEnergy Science Center
Bob Landick, Great Lakes Bioenergy Research Center
Folker Meyer, Argonne National Laboratory
Mary Ann Moran, University of Georgia
Nancy Moran, Yale University
Forest Rohwer, San Diego State University
Doug Rusch, J. Craig Venter Institute
Ramunas Stepanauskas, Bigelow Laboratory for Ocean Sciences
Matt Sullivan, University of Arizona
Tatiana Tatusova, National Center for Biotechnology Information
Niels van der Lelie, Brookhaven National Laboratory

DOE JGI SCIENTIFIC ADVISORY COMMITTEE (SAC)

Bruce Birren (chair), Broad Institute
Toby Bloom, Broad Institute
Jeff Dangl, University of North Carolina
Joe Ecker, Salk Institute
Jim Krupnick, Lawrence Berkeley National Laboratory
Eric J. Mathur, SG Biofuels
Nancy Moran, Yale University
Julian Parkhill, The Sanger Institute
Doug Ray, Pacific Northwest National Laboratory
James Tiedje, Michigan State University
Alexandra Z. Worden, Monterey Bay Aquarium Research Institute

DOE JGI INFORMATICS ADVISORY COMMITTEE (IAC)

Adam Arkin, Lawrence Berkeley National Laboratory

David Dooling, Washington University at St. Louis

Saul Kravitz, MITRE

Stan Letovsky, SynapDx

Jill Mesirov (chair), Broad Institute

Granger Sutton, J. Craig Venter Institute

Kathy Yelick, Lawrence Berkeley National Laboratory

PARTICIPANTS OF WORKSHOP "HIGH PERFORMANCE COMPUTING AND THE NEEDS OF GENOMICS", 2010

Pratul Agarwal, Oak Ridge National Laboratory

Virat Agarwal, IBM

Srinivas Aluru, Iowa State University

Doug Baxter, Pacific Northwest National Laboratory

Pete Beckman, Argonne National Laboratory

Dan Belov, Google

Jay Billings, Oak Ridge National Laboratory

Toby Bloom, Broad Institute

Jim Bristow, Lawrence Berkeley National Laboratory

Jonathan Carter, Lawrence Berkeley National Laboratory

Shane Canon, Lawrence Berkeley National Laboratory

Jarrod Chapman, Lawrence Berkeley National Laboratory/Speaker

Bob Cottingham, Oak Ridge National Laboratory

Tim Davies, Tycrid

Yu Dantong, Brookhaven National Laboratory

Terry Disz, Oak Ridge National Laboratory

Rob Edwards, San Diego State University

Bob Germain, IBM

Rob Gillen, Oak Ridge National Laboratory

Sante Gnerre, Broad Institute

Maya Gokhale, Lawrence Livermore National Laboratory

Richard Goldstein, National Institute for Medical Research, London

David Goodstein, Lawrence Berkeley National Laboratory

Andrey Gorin, Oak Ridge National Laboratory

Chris Henry, Argonne National Laboratory

Phil Hugenholtz, Lawrence Berkeley National Laboratory

David Hysom, Lawrence Livermore National Laboratory

Keith Jackson, Lawrence Berkeley National Laboratory

David Jaffe, Broad Institute

Kirk Jordan, IBM

Jason Ernst, Massachusetts Institute of Technology

Jim Kent, University of California, Santa Cruz

David Konerding, Google

Scott Kohn, Lawrence Livermore National Laboratory

Kostas Mavrommatis, Lawrence Berkeley National Laboratory

Miron Livny, University of Wisconsin

Kamesh Madduri, Lawrence Berkeley National Laboratory

Victor Markowitz, Lawrence Berkeley National Laboratory

Sergei Maslov, Brookhaven National Laboratory

Lee Ann McCue, Pacific Northwest National Laboratory

Folker Meyer, Argonne National Laboratory

Sara Mousa, Google

Cedric Notredame, Center for Genomic Regulation (CRG), Barcelona

Chris Oehmen, Pacific Northwest National Laboratory

Bob Olson, Argonne National Laboratory

Miao-Jung Ou, Oak Ridge National Laboratory

Ross Overbeek, Argonne National Laboratory

Chongle Pan, Oak Ridge National Laboratory

Jignesh Patel, University of Wisconsin

Matteo Pellegrini, University of California, Los Angeles

Len Pennacchio, Lawrence Berkeley National Laboratory

Pavel Pevzner, University of California, San Diego

Jed Pitera, IBM

Phil Papadopoulos, University of California, San Diego

Lavanya Ramakrishnan, Lawrence Berkeley National Laboratory

Dan Rokhsar, University of California, Berkeley

Eddy Rubin, Lawrence Berkeley National Laboratory

Nagiza Samatova, Oak Ridge National Laboratory

Alexander Sczyrba, Lawrence Berkeley National Laboratory

John Shalf, Lawrence Berkeley National Laboratory

David Skinner, Lawrence Berkeley National Laboratory

Tom Slezak, Lawrence Livermore National Laboratory

Bruno Sobral, Virginia Bioinformatics Institute

Alexis Stamatakis, Swiss Federal Institute of Technology, Lausanne (EPFL)

Rick Stevens, Argonne National Laboratory

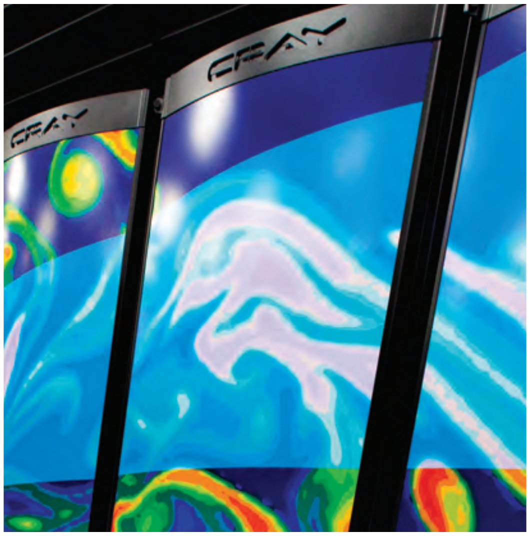
Blair Sullivan, Oak Ridge National Laboratory

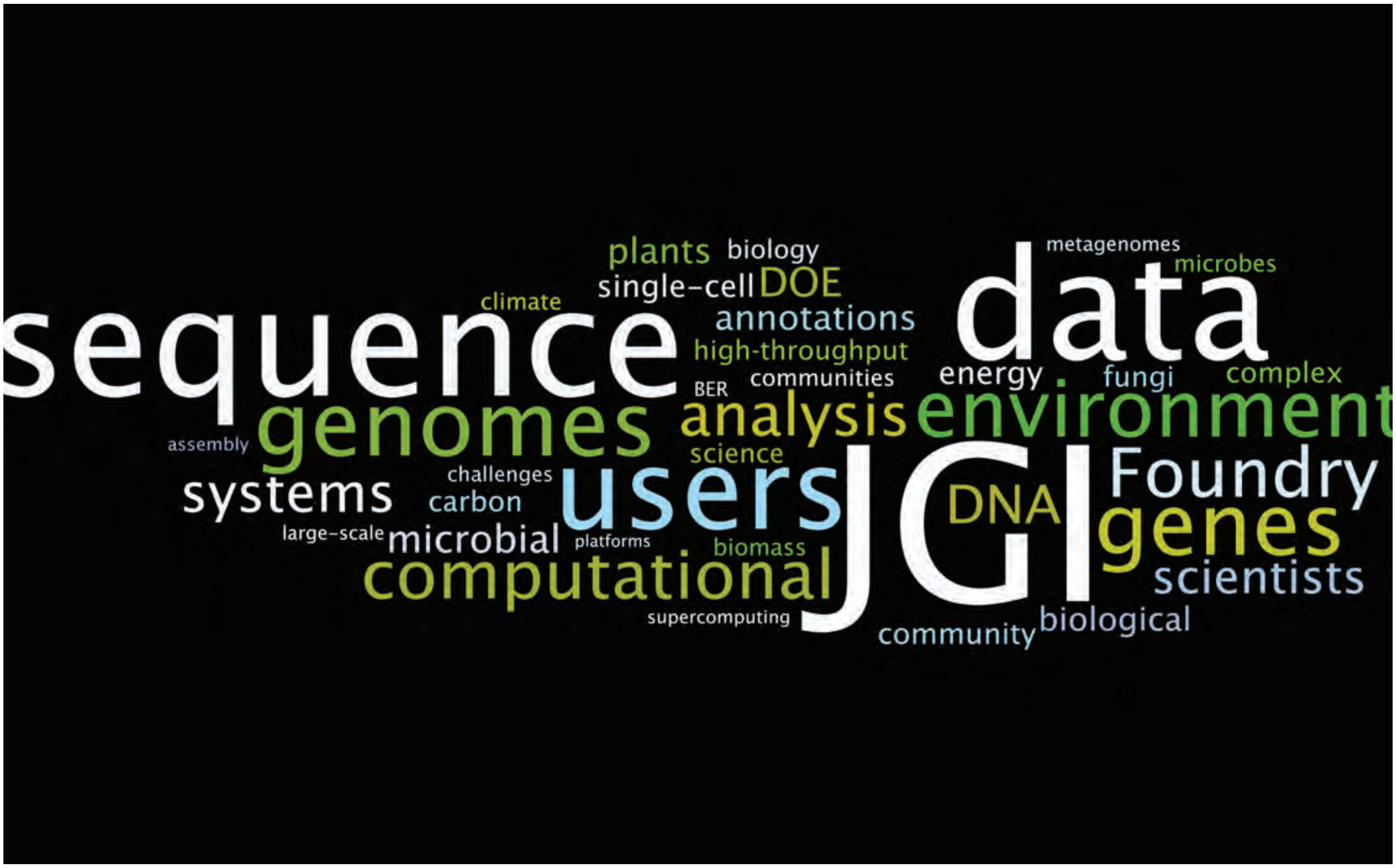
Ward Wheeler, American Museum of Natural History

Fangfang Xia, Argonne National Laboratory

Adam Zemia, Lawrence Livermore National Laboratory


CRAY


CRAY
CRAY



www.jgi.doe.gov

CSO 22655